A new user interface for musical timbre design

Thesis

A new user interface for musical timbre design

Allan Seago BA (Hons) M.Sc

Submitted for the degree of Doctor of Philosophy in Music Computing

September 25th 2009

A new user interface for musical timbre design

Abstract

This thesis characterises and addresses problems and issues associated with the design of intuitive user interfaces for timbral control. The usability of a range of synthesis methods and representative implementations of these methods is assessed, and three interface architectures - *fixed architecture*, *architecture specification* and *direct specification* - are identified. The characteristics of each of these architectures, as well as problems of usability inherent to each of them are discussed; it is argued that none of them provide intuitive tools for the manipulation and control of timbre.

The study examines the nature of timbre and the notion of timbre space; different kinds of timbre space are considered and criteria are proposed for the selection of suitable timbre spaces as vehicles for synthesis.

A number of listening tests, designed to demonstrate the feasibility of subsequent work, were devised and carried out; the results of these tests provide evidence that, where Euclidean distances between sounds located in a given timbre space are reflected in perceptual distances, the ability of subjects to detect relative distances in different parts of the space varies with the perceptual granularity of the space.

Three contrasting timbre spaces conforming to the proposed criteria for use in synthesis are constructed; the purpose of these spaces is to provide an environment for a novel user interaction approach for timbral design which incorporates a search strategy based on weighted centroid localization. Two

prototypes which exemplify the proposed approach in alternative ways are designed, implemented and tested with potential users in order to validate the approach; a third contrasting prototype which represents a simple contrasting alternative is tested for purposes of comparison. The results of these tests are evaluated and discussed, and areas of further work identified.

Acknowledgments

I would like to thank Dr Simon Holland and Dr Paul Mulholland for their supervision and support in the preparation of this thesis.
My thanks also to the students and staff of the Sir John Cass Department of Art, Media and Design at London Metropolitan University and the Music Department at City University, London for participating in the listening tests.
In particular, I would like to thank Kendall Wrightson, Christina Paine and Lewis Jones for their support and comments in the final stages of the thesis.
Finally, my thanks to Karen for her love and support throughout.

PortAudio RB, used in the test software, is an open source library based on PortAudio by Bencina and Burk (2004).

Publications

The HCI perspective on sound synthesis hardware and software in chapter two has appeared in Seago, A., S. Holland, P. Mulholland (2004). A Critical Analysis of Synthesizer User Interfaces for Timbre. HCI 2004: Design for Life, Leeds, British HCI Group.

Some parts of the perceptual study of a simple timbre space described in chapter five have appeared in Seago, A., S. Holland, P. Mulholland (2005). Towards a Mapping of Timbral Space. Conference on Interdisciplinary Musicology (CIM05), Montreal, Canada.

Preliminary results from the testing of the weighted centroid localization strategy described in chapters six and seven appear in Seago, A., S. Holland, P. Mulholland (2008). Timbre space as synthesis space: towards a navigation based approach to timbre specification. Institute of Acoustics Spring Conference 2008: Widening Horizons in Acoustics, University of Reading.

Table of Contents

Table of Figures

CD-ROM

I. Raw data from chapter two task analysis.

II. Web pages used in listening tests in chapter five.

III. Software used in chapters six and seven: executable files (Mac OS X 10.5)

IV. Source code

Chapter 1 - Introduction

1.1. Motivation

The appearance of the first musical instruments to use purely electronic methods to generate sound can be dated variously to Elisha Gray's *Musical Telegraph* of 1874 and Thaddeus Cahill's *Telharmonium* of 1897. Since then, a wide range of synthesis techniques have become available to musicians, composers and other practitioners. Some of these, such as additive synthesis, offer direct low-level control over acoustical attributes of sound – the overall spectral envelope, or the amplitude of individual spectral components, for example. Others are based on a concept of sound as the output of a network of functional components, which might variously be oscillators and filters in the case of classic voltage control synthesis, or simulations in software of acoustical components such as pipes, strings, membranes and plates. A third category is the family of synthesis methods which are implementations of some mathematical abstraction, such as frequency or amplitude modulation, and whose parameters do not easily map to specific sonic attributes.

The synthesis paradigm employed in the early years of the twentieth century for the electronic generation of sound was, for the most part, that of *additive synthesis* – the building of complex spectra from a number of harmonic or inharmonic sinusoidal components. Cahill's *Telharmonium*, for example, generated sounds by use of tonewheels, each one associated with a harmonic (Weidenaar, 1995). One of the earliest instruments to use *subtractive synthesis*, which is the selective filtering of complex spectra to produce the desired sound, and which can be seen as the inverse of additive methods, was Friedrich Trautwein's *Trautonium*

of around 1929; this method of synthesis formed the basis of the voltage controlled synthesizers of the 1960s and 1970s.

The rapid development of digital technologies in the following decades brought in its wake a considerable expansion in the range of techniques available to the composer of electroacoustic music. Frequency modulation (Chowning, 1973) is a technique adapted from radio communications which allows the generation of complex sound spectra by modulating, at audio frequency rates, the frequency of a carrier oscillator. In waveshaping synthesis (Arfib, 1979), sounds are generated using a transfer function to map an input waveform (typically a sine wave) to the desired output. Granular synthesis (Gabor, 1947; Xenakis, 1971; Roads, 1988; Truax, 1988) can be used to create complex sound events from the accretion of thousands of sound 'grains', each having a duration of the order of milliseconds. Wavetable synthesis signals are generated by repeated digital-to-analogue conversion of a table of values representing one cycle of the waveform. Formant synthesis (Rodet, 1984) makes use of formant frequencies and amplitudes characteristic of human voices and musical instrumental sounds. Physical modelling synthesis (Smith, 1992) is based on the simulation in software of acoustical mechanisms, such as those of vibrating strings, pipes and plates, which can then be 'struck', 'plucked' or 'blown'.

While the range of tools and techniques available to musicians, sound designers and composers for the design and editing of sound is thus very wide, usability - the ease with which a task can be completed or a goal achieved using a particular tool or system – in both hardware and modern software synthesizers is generally poor (Ethington and Punch, 1994; Miranda, 1995; Rolland and Pachet, 1996). Usability has been defined by a number of components: learnability, efficiency, memorability, user satisfaction, the number of errors that a user makes,

and the provision of means to recover from errors (Nielsen, 1994). None of these features is strongly characteristic of the modern synthesizer interface. This has led to a situation where most synthesizer users seem to have limited their choices of timbre to selections from a bank of preset timbres - evidence for this is largely anecdotal, but 'allegedly, nine out of ten DX7s coming into workshops for servicing still had their factory presets intact' (Computer Music, 2004) .

Synthesis techniques were originally implemented in hardware synthesizers; however, over the last decade, this functionality has increasingly been migrating into software (*Reaktor*[1], *Reason*[2] etc). This development has potentially freed designers from the constraints imposed by hardware limitations, particularly from the limited space available for controllers and displays, but also from cost constraints of hardware controls. Yet software designers have sought to emulate hardware synthesizers, not only in models of synthesis – how the sounds are generated - but also in the user interface. Thus, the user is presented on screen with a simulation of a legacy synthesizer hardware control surface, and must control it via virtual buttons, faders and rotary dials that mimic the hardware they have replaced. For many users, this has the virtue of familiarity, but at the same time, means that issues of usability (the ease with which sounds can be created and edited) which arise in the design of synthesizer hardware apply with equal force to their software equivalents.

In other software application domains – word processing, graphic design, music notation, financial etc. - the trend over the last fifty years has seen the user/system boundary shift towards the user; that is to say, the interaction which the user has with the system has increasingly been expressed in user-oriented, rather than system-oriented terminology. Sound synthesis technology, by contrast,

[1] Native Instruments
[2] Propellerhead Software

has not undergone this process to the same extent, and the user is obliged to express directives using the terminology of the particular synthesis method in use. Thus, the successful negotiation of the interface of a commercially available synthesizer depends to a great extent on, firstly, the domain knowledge that the user brings to the task, and, secondly, the degree of practical experience that the user has of similarly configured instruments (Seago, 2004). A user without this knowledge and experience will be deterred from tackling anything other than the most basic editing tasks.

The prevalence of the subtractive synthesis paradigm in the design of commercially available synthesizers means that the terminology associated with it seems to have become a *lingua franca* for sound synthesis in general. This became apparent during user testing (discussed later in this thesis) of the interfaces of two commercially available synthesizers (the Roland XP-50 and the Korg Trinity) (Seago, 2004). Many of the subjects were broadly familiar with simple four-stage envelope generation such as attack, sustain, decay and release (ADSR) for the control of loudness[3] and filtering for controlling timbre, and expected to find these reflected in the interface. A level of expertise in the principles of subtractive modular synthesis, and its associated terminology is thus, to some extent, transferable between different models and makes of instrument. However, for musicians who are not conversant with the language of oscillators, filters, and voltage control, and who would prefer to work with more obviously musical terminology and concepts (terms such as *brightness*, *openness*, *compactness* and *acuteness* for example), the process of creating and editing sounds on hardware and software synthesizers can be daunting.

[3] It should be noted that ADSR is not uniquely associated with subtractive synthesis; however, it has been an important feature of subtractive synthesizers since the late 1960s.

There exists a considerable body of research, reported in such fora as NIME[4], which is concerned with the development of novel means of interacting with sound-generating technology in real-time, for the purposes of musical performance. This is not, however, the subject of this thesis. While the NIME community also engages with many of the issues and concerns raised here, it is not our intention to analyse or to consider the usability of interfaces intended for extended 'live' performance. Instead, our focus is on the usability problems inherent in modern commercially available synthesizers, and on the means of specifying the characteristics of a single discrete sound object. Such usability problems have received relatively little attention in the field of human-computer interaction (HCI), and an important part of this thesis is to review that research which exists, and to conduct an analysis, from the HCI perspective, of the means of shaping and editing timbre in a number of typical modern hardware and software synthesizers.

1.2. Aims and objectives

The overall research question to be addressed in this thesis is the extent to which timbre space, a construct that has been successfully used as an analysis model for understanding the psychoacoustical basis of timbre perception, can be used as a search space for sound synthesis. Specifically, the aims of the thesis are:

- to propose and describe a novel timbre space search strategy for timbral shaping, in which iterative user input drives a software process where a candidate solution converges on the desired sound,

[4] New Interfaces for Musical Expression

- to build a number of software prototypes which embody different versions of this strategy in a number of different timbre spaces,

- to present the findings of a series of user tests of the prototypes and

- to demonstrate that the strategy has potential as the basis of a user interface for navigating timbre spaces whose dimensions are 'well-behaved' and represent acoustical attributes which are perceptually linear.

In order to support and contextualise the discussion and to provide a rationale for the proposed search strategy, the thesis has the following subsidiary objectives.

The first of these is to characterise the problems and issues associated with the design of intuitive user interfaces for timbral control. The user-system dialogue in a number of representative current hardware and software synthesizers is investigated and analysed from a usability perspective, with particular attention given to the controls available to the user for specifying and controlling musical timbre. From this discussion, a taxonomy of three interaction styles for manipulating timbre in synthesizers is proposed, and the synthesis methods most suited to each of these styles identified. A number of hardware and software synthesis implementations representative of these three styles are analyzed.

The second objective is to explore the notion of timbre space, to identify and describe candidate timbre spaces suitable for search, and to propose a set of criteria for an ideal *n*-dimensional attribute space which functions usefully as a vehicle for sound synthesis.

Lastly, and in order to provide evidence that the chosen timbre space is navigable by the search strategy, it is intended to demonstrate the extent to which Euclidean distances between sounds disposed in the space are reflected in perceptual distances; this will be achieved by the running of a number of listening tests.

These three subsidiary objectives serve the overall aim described above; namely, the design and testing of a prototype search strategy for timbral shaping.

1.3. Structure of the thesis

Chapters two to four inclusive comprise the literature review of the thesis. The interdisciplinary nature of the work presented here means that relevant work from the HCI perspective (chapter two), in sound and music perception (chapter three) and music computing (chapter four) is reviewed.

Chapter two begins by examining a selection of commonly used synthesis methods, from the point of view of usability. This part of the discussion draws on and discusses an established taxonomy of synthesis techniques (Smith, 1991), together with a set of evaluation criteria proposed by Jaffe (1995).

The discussion then moves on to consider the current implementation of synthesis methods both in hardware and software. Beginning with a review of HCI work in this area, it goes on to consider a model of user/system interaction which can be applied to the specification of sound, drawing on established HCI terminology in order to frame the problem as one of human-computer interaction. Three distinct synthesis architectures are identified and defined in this chapter: *fixed architecture* (where parameter values are specified through menus and/or

forms), *architecture specification* (in which a sound is conceived as emerging from a user specified network of functional components), and *direct specification* (where the user engages with some visual representation of the sound). Each of these is analysed from the point of view of usability, and synthesis methods appropriate to each architecture identified. Chapter two ends by addressing the question of why direct specification methods are not universally used for sound synthesis, or at least for those synthesis techniques that are identified as being best suited to direct specification.

This question is addressed in the study of timbre contained in chapter three, which begins by noting issues of terminology and definition. Such issues are, in themselves, symptomatic of the wider problem; the well known ANSI definition (American National Standards Institute, 1973), for example, is essentially negative, in that it defines timbre in terms of what it is not; and writers on this topic have variously considered timbre as the instantaneous colour of the sound, or more globally, as the overall characteristics of the sound as it evolves over time, or (more globally still) as a perceptual phenomenon which subsumes musical parameters such as pitch and loudness.

Because this thesis is concerned with the design of tools for timbral articulation in electronic musical instruments, chapter three briefly considers the role of timbre in music of the nineteenth and twentieth centuries. The foregrounding of the timbral element of music during this period has not resulted in any generally accepted theory of musical timbre, possibly because timbre (unlike pitch, loudness and rhythm) does not lend itself to description using discrete symbols or scalar values. Those that exist, however, are reviewed in this chapter, together with a contrasting approach to the theorising of timbre in which

sound is seen as arising from the interaction of physical processes within physical environments – the 'ecological' perspective.

However, while such theories provide useful theoretical underpinnings for the artistic use of timbre in soundscapes and musical structures, they are less helpful when it comes to providing tools for the control of timbre which are sufficiently specific, while at the same time being both flexible and intuitive (in the sense that the steps needed to create and to edit a timbre are clear and obvious); and it is for this reason that the main focus of chapter three is on timbre as a psychoacoustical phenomenon, and specifically on research that has sought to identify those acoustical correlates which contribute significantly to our perception of timbre.

This section of chapter three begins with a discussion of the spectral component of sound and its association with timbre, before going on to consider the importance of its overall dynamic envelope – attack and decay characteristics. A distinction is made here between *timbre perception* – what acoustical attributes are salient to our perception of timbral change - and *timbre identification* – what acoustical attributes govern our ability to identify the source of a sound. The importance of this distinction is that a sound synthesis system might be variously used to replicate the sound of an existing musical instrument, to create a hybrid sound whose attributes are those of two or more existing musical instruments, or to design a completely new sound whose qualities are not those of any existing acoustical source.

An alternative approach to the understanding of timbre has been to investigate correspondences and mappings between measurable acoustical parameters of sound and the rich lexicon of descriptive terms available to

musicians for its verbal description. Such connections have implications for the design of user interfaces for timbre and for this reason, research methodologies and findings in this area are reviewed in this chapter, and their applicability to user interface design assessed.

Lastly, and most importantly for this thesis, chapter three considers the notion of *timbre space*. Because timbre, unlike pitch and loudness, is multidimensional, a useful model is that of an *n*-dimensional coordinate space, each of whose dimensions is a vector representing some variable acoustical attribute. Again, a distinction is made between, firstly, those timbre spaces whose dimensions are predetermined (*attribute spaces*) and, secondly, those which are constructed from data derived from listening tests, and the nature of whose axes are only determined subsequently (*perceptual spaces*). (For the remainder of this chapter, however, the term timbre space will be retained.) Of particular relevance to the empirical work presented in this thesis is the extent to which one type of space can be mapped to the other – specifically, whether relative Euclidean distance relationships between sounds in a given space are reflected in perceptual distances. Two papers (Ehresman and Wessel, 1978; McAdams and Cunible, 1992) which investigated such relationships are discussed in this section of chapter three.

An important technique that has been used for the construction of timbre spaces (and is applied in part of the empirical work presented in this thesis) has been multidimensional scaling (MDS). Chapter three concludes by reviewing important studies based on this technique, with particular reference to the work of Caclin, McAdams, Smith and Winsberg (2005), whose timbre space was based on the findings of a number of MDS analyses, and which forms the basis of one of the spaces used in chapter six.

Chapter four returns to the topic of sound synthesis systems, but this time specifically to review applications which have sought to bridge the gap between user perception of timbre and the parameters of synthesis. The drawbacks of direct specification methods using graphical user interfaces, for example, have been addressed in two papers, reviewed in this chapter, which propose interfaces based on synthaesthetic links between colour and texture. Most of the chapter, however, is concerned with work that builds on the timbre space model and/or which applies artificial intelligence (AI) techniques to the specification of sound. A usable and sufficiently flexible system based on timbre space needs to address the problem of timbral multidimensionality; the work reviewed in this chapter frames this problem as one of data reduction, either using principal component analysis (PCA) or some form of MDS. The work of Hourdin, Charbonneau and Moussa (1997), is highlighted here and described in detail; they proposed a 'musical space' of reduced dimensionality which was derived from a number of orchestral instrument samples using MDS, and which forms the basis of the search space described in chapter seven of this thesis. Other researchers have applied techniques drawn from the field of artificial intelligence (AI). Two distinct approaches are identified and reviewed here (although it should be noted that a number of studies incorporate elements of both).

In the first, the process of specifying and editing a sound is defined as a knowledge-based activity, and accordingly makes use of knowledge-based systems (KBS), either to apply encoded rules and heuristics relating to synthesis expertise, or to map synthesis parameters to the adjectives and adverbs used to describe sound. The advantages and drawbacks of this approach will be considered here.

The second approach treats the process as one of evolutionary search. Algorithms which implement evolutionary search strategies are discussed here, with particular attention given to genetic algorithms. These are evolutionary search strategies inspired by selection mechanisms occurring in nature, such as inheritance, mutation and crossover, to drive a process of optimisation, in which the individuals of a population are iteratively selected according to some predefined fitness function. A particular subset of this approach is considered here, in which the fitness of individuals is determined by the user – so-called 'interactive genetic algorithms'. A study which is foregrounded here is the work of McDermott, Griffith and O'Neill (2007), who proposed a system to address the 'bottleneck' which is a characteristic of such systems. Because there are a number of similarities, as well as important differences, between this work and the work presented in this thesis, this paper will be considered in detail, and the conclusion of the thesis will draw broad comparisons between the two.

Chapter five introduces the empirical work. As stated earlier on, its aim is to assess the operation of a weighted centroid localisation strategy driven by similarity-dissimilarity judgments, based on iterative updating of an *n*-dimensional probability table, in three distinctly different timbre spaces. However, its success is entirely dependent on the ability of the user to perceive relative Euclidean distances in the space. This chapter discusses the methodology and results of a series of listening tests. The tests are designed to determine whether in general, subjects, when presented with three sounds A, B and C, disposed within a previously constructed three-dimensional attribute space, such that the distance AC is greater than the distance BC, are able to perceive those relative distances as degrees of timbral difference. The study shows that subjects are, in fact, able to do this with an average accuracy of 73.02%, and concludes that this particular space is a suitable vehicle for testing of the search strategy.

Chapter six introduces the timbre space search strategy and discusses its operation in two contrasting three-dimensional timbre spaces. The search strategy employs an adapted *weighted centroid localisation* (WCL) algorithm, which is used to drive the convergence of a candidate search solution on to a 'best-fit' solution, based on user input. One of the two spaces in which the strategy is tested is that used in the listening tests of chapter five, and which we will call the *formant space* (for reasons explained later). The other one, called the *SCG-EHA space*, (again, for reasons explained later) is derived from the MDS space generated by Caclin *et al* (2005); it is argued here that, because the mapping in this space between physical and perceptual dimensions has been shown to be robust (the purpose of the Caclin *et al* study was to demonstrate this), it too is a suitable vehicle for testing of the search strategy, and that no listening tests of the type described in chapter five need be conducted. The characteristics of the SCG-EHA spaces are described in detail (those of the formant space are described in chapter five).

Having considered the two timbre spaces, the second section of chapter six then describes the WCL algorithm in detail. The third section describes the testing procedure, in which the operation of three search strategies in the two timbre spaces was evaluated. Two of the three strategies were variants of the WCL algorithm. The third strategy, however, was not a WCL implementation; instead, it provided the subject with an interface which afforded direct access to the axes of the space being investigated, in the form of sliders. Adjustment of the position of each slider altered the value of the corresponding parameter. We have called this the *multidimensional line search* (MLS) strategy; the purpose of its inclusion was to determine whether the WCL strategy (in either form) performed significantly better than an MLS interface which provided direct access to the parameters. Results from these tests showed that both the WCL strategies performed

significantly better than the MLS strategy.

The two timbre spaces investigated in chapter six were of low dimensionality and limited coverage. Chapter seven considers whether these results can be generalised to more complex and 'realistic' timbre spaces, and describes in detail the process of constructing a seven dimensional MDS timbre space from heterodyne analysis of a selection of samples of orchestral instruments. Both the characteristics of this space and the means by which it was derived – MDS - were similar to those of the space constructed by Hourdin et al (1997). It is demonstrated that the frequency spectra of the sounds in the MDS space are comparable to those of the original sounds.

The remainder of chapter seven is devoted to describing the testing of the WCL and MLS strategies in this MDS space, and to reviewing the results. Again, it was found that the WCL strategies performed better than the MLS strategy.

The final chapter summarises the research, considers synthesis engines appropriate to the WCL strategy, proposes ways in which the WCL strategy could be realised in a practical implementation, and outlines some directions for further work.

Chapter 2 - Sound synthesis and the synthesizer interface

2.1. Introduction

This chapter reviews and examines sound synthesis from the point of view of usability. The first part describes a number of synthesis techniques themselves and their very diverse parameters and employs a taxonomy drawn from Smith (1991); we note that, while the parameters of some link very well to audible features of sound, this is less obviously the case with others. The discussion then moves on to consider a number of criteria developed by Jaffe (1995) for assessing synthesis techniques; the section closes by describing the methodology and results of an evaluation that draws on these criteria (Tolonen, Välimäki *et al.*, 1998).

Synthesis methods themselves can be viewed as abstractions and as such are not easily susceptible to usability evaluation methods employed in other application domains; only where they have been implemented in hardware or software can such methods be usefully applied. The second part of this chapter considers the usability of a number of representative synthesis implementations. Human-computer interaction (HCI) research has been applied in recent years to novel methods of controlling sound in real-time (Vertegaal, 1994; Vertegaal and Eaglestone, 1996; Hunt, Wanderley *et al.*, 2000; Wanderley and Orio, 2002; Wessel and Wright, 2002). However, the user interfaces of audio hardware and software generally, and of music synthesizers in particular, as distinct from performance based 'real time' interfaces, have received relatively little study within the field of

HCI (Ruffner and Coker, 1990; Polfreman and Sapsford-Francis, 1995; Fernandes and Holmes, 2002).

Thus, this second part of the chapter opens with a brief review of HCI oriented work which exists in this area. It is followed by a discussion of a model of interaction drawn from Dix, Finlay, Abowd and Beale (1998), and considers the extent to which it can be generally applied to the specification of sound. In particular, terms such as *task language* and *core language* are introduced and defined, and their use explored in the context of sound synthesis.

The discussion then moves on to a consideration of a number of representative synthesis implementations. Three distinctive architectures are identified and defined, and a heuristic evaluation conducted on each, in order to highlight key themes which will be addressed throughout the thesis.

2.2. The synthesizer user

Discussion of the user-system dialogue in other domains normally begins by identifying who the users are (Dix, Finlay *et al.*, 1998). In the case of the standard commercially available hardware and software synthesizer, however, this is not straightforward, as there is no typical user. They may be hobbyists or professionals, performers, researchers or composers; they may be working in commercial and mainstream spheres, or composers and performers of avant-garde music. The intended uses of the synthesizer may be for music and sound design in film, television and video, for dance, rock or for the classical concert hall. As noted in the introduction, users may explore the timbral possibilities offered by the synthesizer interface and the synthesis engine which sits behind it; more often than not, however, use of the

instrument is confined to the available preset sounds provided by the manufacturer. Polfreman and Francis (1995) conclude (as have other researchers) that this limited use of the technology is largely because of the poor usability of the tools provided for the invention of new sounds and the modification of existing ones.

Because one of the aims of this thesis is to examine and discuss these issues, the user will be considered here to be one whose use of the synthesizer may fall into any one of the categories (performer, researcher, composer etc) described above, but whose interests and concerns, nevertheless, extend beyond the passive use of presets to the active exploration of the technology in order to devise new and interesting sounds.

2.3. The user interface

The user interface is the set of software and/or hardware components which together facilitates the means of communication between a human and a machine or system. In the case of modern commercial hardware synthesizers, the controls for specifying and editing timbre typically (but not exclusively) consist of rotary dials, buttons and sliders for specifying synthesis parameters, and LCD displays for indicating the current state of the edited sound.

Before considering the synthesizer user interface in greater detail, we examine the various parameters of sound synthesis engines.

2.4. The parameters of sound synthesis

Control parameters for currently available methods of sound synthesis vary considerably. Some of them map very readily to perceived timbral qualities (a rise in the cut-off frequency of a low pass filter, for example, is heard as an increase in the 'brightness' of a tone). Others, such as those of frequency modulation (FM) and amplitude modulation (AM) provide very little audible association with any single sonic attribute; such synthesis methods have been described as 'loose modelling' approaches for this reason (Miranda, 2002) . Techniques such as additive synthesis require a considerable number of low level control parameters (e.g., the amplitudes of individual harmonics), whereas the parameters of other techniques are fewer in number and are more high level; a change in the values of the 'mass' or 'stiffness' parameters associated with physical modelling will cause changes in a number of lower level aspects of the generated sound, such as decay time, spectral width etc.

The following sections 2.4.1 to 2.4.4 list and briefly define and tabulate a number of established digital synthesis methods, each of which presents the user with a distinct set of control parameters at various levels of abstraction. They are grouped according to a taxonomy based on (Smith, 1991), in which four categories of synthesis are identified - *abstract algorithms, spectral models*, *sampling* or *processed recordings* and, lastly, *physical models.* The list of synthesis methods is by no means exhaustive, but is based on one compiled in an evaluation study (Tolonen, Välimäki *et al.*, 1998) (to be discussed later) and broadly covers current approaches to synthesis. It should also be noted that the categorisation is not absolute, and a synthesis method may fall into more than one category.

2.4.1. Abstract algorithms

These are synthesis techniques based on methods that may be explorations of a mathematical expression, but have little to do with real-world sound production mechanisms or with perceived attributes of sound. *Frequency modulation synthesis*, for example, is inspired by FM radio transmission (Chowning, 1973), in which the frequency of a carrier signal is controlled by the frequency and amplitude of a modulator signal, creating sidebands around the carrier frequency resulting in a complex spectrum. *Waveshaping* is based on the principle of non-linear distortion of a sine wave; the frequency content of the complex spectra generated is harmonic and related to the transfer function of the distortion process. Useful transfer functions can be constructed using weighted combinations of Chebyshev polynomials; a fourth-order Chebyshev polynomial, for example, will output a sinusoid whose frequency is four times that of the input. *Karplus-Strong synthesis* makes use of a filtered digital delay line; the output from the delay is fed back into the input, resulting in a progressive attenuation of higher harmonics; because it can effectively simulate the vibrations of a plucked or hammered string, it can also be seen as a special case of physical modelling using digital waveguide synthesis (described in section 2.4.4).

The following table summarises the above methods, together with the parameters normally associated with them.

Method	Indicative parameters
Frequency modulation (FM)	Carrier/modulator frequency ratio and modulation index.
Waveshaping	Number and order of Chebyshev polynomials, weighting of Chebyshev polynomials
Karplus-Strong (KS)	Size of lookup table

Figure 2.1: Synthesis methods based on abstract algorithms, and their control parameters.

2.4.2. Processed recording

Synthesis methods in this category take existing sound events and either reproduce them directly or process them further to create new sounds. Of these three, *sampling* is the simplest, employing a table containing a digital recording of usually no more than a few seconds in length, which can be looped and pitch shifted. The only other means of control is at sample level – that is to say, control over the value of individual samples. Simple *wavetable* synthesis works in a similar way, except that, typically, only one cycle of the desired waveform is stored. *Multiple wavetable* methods, making use of more than one wavetable, are essentially an extension of this; the wavetable used for the onset or attack of the note is cross-faded with the wavetable used for the decay, which is, in turn, cross-faded with that used for the sustain phase, and so on. In *granular synthesis*, sequences or bursts of very small 'atoms' or 'grains' of sound are generated, either sequentially or scattered over the time-frequency plane in some pre-specified pattern of distribution, producing complex time-variant sounds.

Method	Indicative parameters
Sampling	Loop length, pitch shift, individual samples
Multiple wavetable	Attack, decay sustain, release
Granular synthesis	Grain envelope, duration, shape, waveform and frequency. Delay time between grains

Figure 2.2: Synthesis methods based on processed recording, and their control parameters.

2.4.3. Spectral models

These are synthesis techniques which afford manipulation of the spectral properties of sound as it is perceived by the listener. Sounds created using *additive*

synthesis are generated by a summing of sinusoidal components according to the Fourier theorem. An effective synthesis using this method requires a considerable amount of control data and is computationally expensive. *Subtractive synthesis* (referred to as *source-filter synthesis* in Tolonen *et al (*1998*)*) is perhaps the most well known of all current sound synthesis methods, forming the basis of the voltage controlled synthesizers of the 1960s and 1970s, and involving the filtering of a spectrally rich waveform in order to obtain the required output. In some respects, it has features in common with *physical modelling synthesis*, in that sound is viewed as the output of a network of functional components – oscillator, filter, amplifier etc. *Formant synthesis* - the example here uses the *fonction d'onde formantique* (FOF) technique (Rodet, 1984) - is based on the premise that the spectra of many vocal and instrumental sounds are characterized by distinctive peaks called formants (see chapter three for a discussion of formants and their importance in timbre perception). The impulse responses of a set of filters, each of which corresponds to a formant, are derived from analyzing a recorded signal; these filters are then used for synthesis.

Both Smith (1991) and Tolonen *et al* (1998) include VOSIM (VOice SIMulation) in this list because it is seen as a variant of formant synthesis. However, because this synthesis method uses bursts of pulses of decreasing amplitude, and its parameters define (amongst other things) the number and duration of those pulses, and the delay time between them, Tolonen *et al* note that this method can also be regarded as a form of granular synthesis.

Method	Indicative parameters
Additive	Amplitudes of individual harmonics
Subtractive (source-filter)	Cut-off frequency, Q (resonance), centre frequency
Formant synthesis (e.g., FOF)	Formant centre frequency Formant amplitude Rise time of local envelope (bandwidth of formant at –40 dB) Decay time of local envelope (bandwidth of formant at –6 dB)
VOSIM	Number and duration of pulses, delay between pulses Initial amplitude Multiplying factor

Figure 2.3: Synthesis methods based on spectral models, and their control parameters.

2.4.4. Physical models

These are synthesis techniques which seek to simulate the sound production mechanisms of real world musical instruments, the central idea being "to start with the known and then extend it in some direction" (composer David Jaffe, quoted in Chadabe (1997)). One version of this is *modal synthesis*, which operates at quite a high level of abstraction, modelling both the material properties (mass, stiffness, tension etc) of acoustical components such as tubes, membranes and soundboards etc, and the types of interaction that are possible between them. *Digital waveguide synthesis* employs computational models of wave propagation through material media, making use of delay lines (as in KS synthesis) to represent the geometry of the medium.

Method	Indicative parameters
Modal synthesis	Physical properties of simulated real world objects – mass, tension, stiffness etc Nature of interaction – striking, plucking etc
Digital waveguide	Size of delay line, gain factor etc

Figure 2.4: Synthesis methods based on physical models, and their control parameters.

2.5. Existing evaluations of synthesis methods

Evaluative analyses of current synthesis methods, certainly from the usability perspective, are scarce. However, a report produced by Helsinki University of Technology (Tolonen, Välimäki *et al.*, 1998) reviewed this list of digital synthesis methods, based on Smith's taxonomy, and which used a set of criteria proposed by Jaffe. Both these criteria (Jaffe, 1995) and the evaluation based on them will now be discussed.

Jaffe's ten criteria relate to three main areas, only two of which are of interest to the subject of this thesis; the usability of the parameters and the range and quality of sounds produced. (The last of these categories has to do with issues such as latency, memory usage and processing requirements and will not be considered here.) Firstly, Jaffe states that there should be a clear and predictable link between a given parameter and an audible sonic attribute. Some methods of synthesis map very readily to perceived timbral qualities (a rise in the cut-off frequency of a low pass filter is heard as an increase in the 'brightness' of a tone, for example). Associated with this is the requirement that a parameter should be 'well-behaved'; a degree of change in parameter value should produce a proportional perceptual change in the sound. The parameters of 'loose modelling approaches' such as frequency modulation (FM) and amplitude modulation (AM) synthesis (Miranda, 2002) present usability problems in this regard; small changes to the modulator frequency or amplitude in a modulator-carrier pair can result in considerable and (from the user's perspective) unpredictable timbral change.

Secondly, parameters should be 'powerful', in that any change in value should have an audible effect. Again, a change to the cut-off frequency of a filter will cause a very obvious change to the timbre of the sound, whereas changing the amplitude of a single harmonic in additive synthesis is less perceptible. (Jaffe notes, however, that a set of 'weak' parameters can be grouped together to form a 'stronger' one – a 'metaparameter'.)

Thirdly: how extensive is the coverage of a given synthesis technique? Is there a class of sound that it cannot generate? Pure wavetable synthesis and heterodyne analysis-resynthesis (a form of additive synthesis) cannot produce inharmonic sounds, whereas time-varying additive synthesis, in general, has the potential (in theory) to produce almost any sound, given sufficient control data.

Fourthly: are there well-understood analysis techniques by which synthesis parameters can be derived from real world models? Acquiring the correct carrier/modulator frequency and amplitude settings for the imitative recreation of a given sound is extremely difficult (chapter four will identify and discuss AI approaches to this particular problem).

Finally, Jaffe proposes that a synthesis method should generate sounds which retain identity in the context of variation. Discussion of what constitutes invariance in timbral perception and the distinction to be made between timbre perception and timbre identification is deferred to the next chapter; however, the proposal raises the question of the extent to which a given sound can be attributed to a physical source or be described by some commonly understood adjective or adverb.

Jaffe's criteria provide a useful set of heuristics for evaluating the usability of different synthesis techniques, and were used by Tolonen *et al* as the basis of their examination of the range of synthesis techniques, some of which are listed above. The methodology used by Tolonen *et al* was to evaluate each technique according to the areas of interest proposed by Jaffe – *usability*, *quality and diversity of produced sounds* and *implementation*. (Again, for the purposes of this discussion, implementation will not be considered.)[5]

Within the general rubric of usability are a number of sub-criteria - *intuitivity*, *perceptability*, *physicality* and *behaviour*. *Intuitivity* is defined here as the extent to which 'a control parameter maps to a musical attribute or quality of timbre in an intuitive manner' (p. 86), and the ease in which a user might learn how to control the synthesis engine. *Perceptability* relates to the strength or weakness of the control.- whether it causes clearly audible changes. *Physicality* is the extent to which the parameters relate to the behaviour of 'real world' acoustical mechanisms. *Behaviour* is a measure of whether small or large changes in a parameter value are reflected in proportionally small or large changes in the quality of the sound.

The quality and diversity of the produced sound is considered under Jaffe's indicators of *robustness of identity*, *generality* and the availability of *analysis methods* to drive the synthesis. *Robustness,* in this context, means the extent to which the sound retains its identity when parameter values are varied; for example, a synthesised 'clarinet' should still sound recognisably like a clarinet when parameters which affect or otherwise relate to dynamics or blowing styles are

[5] While the range and quality of the generated sound are not in themselves the focus of this thesis, clearly there is a trade off between these factors and usability. A simple, intuitive interface is of little use if the range of available sounds is very restricted. Barry Truax's discussion of the relationship between 'generality' and 'strength' is relevant here. Truax, B. (1980) The inverse relation between generality and strength in computer music programs. *Journal of New Music Research*, 9(1), 49-57.

modified. (Again, discussion of the issues of timbral identity which this raises is deferred to the next chapter.) *Generality* is a measure of the synthesis method's coverage – the range of sounds it can produce. Finally, many synthesis methods require a previous 'offline' *analysis* of audio samples in order to derive synthesis parameter values; the evaluation performed by Tolonen *et al* considered the availability, accuracy and generality of such methods, as well as the extent to which specialised techniques or instruments are required.

Not all the criteria were applied to all the methods surveyed in the work of Tolonen *et al*. For example, sampling is controlled simply by 'start', 'finish' and gain parameters (which are trivial). Similarly, multiple wavetable synthesis methods can be parameterised in a number of ways and the result of synthesis is highly dependent on the nature of the signals stored in the wavetables themselves. In these cases (and some others), no rating was given. The report does not indicate whether any preferred structured evaluation method was employed – whether, for example, a task analysis was performed which could form the basis of a cognitive walkthrough (Wharton, Rieman *et al.*, 1994). The results seem to have been arrived at through a fairly subjective methodology; each method was rated on a scale of one to three (one being *poor*, two *fair* and three *good*) for each criterion.

The findings of Tolenen *et al* are summarised below. For the evaluation of formant synthesis, the CHANT system was used (Rodet, Potard *et al.*, 1984). Figure 2.5 ranks a number of synthesis methods in order of parameter usability (note that sampling and multiple wavetable do not appear here for the reason given above) – the ratings from each sub-category are summed; figure 2.6 shows the same data broken down into the sub-categories of *intuitivity*, *perceptibility*, *physicality* and *behaviour*.

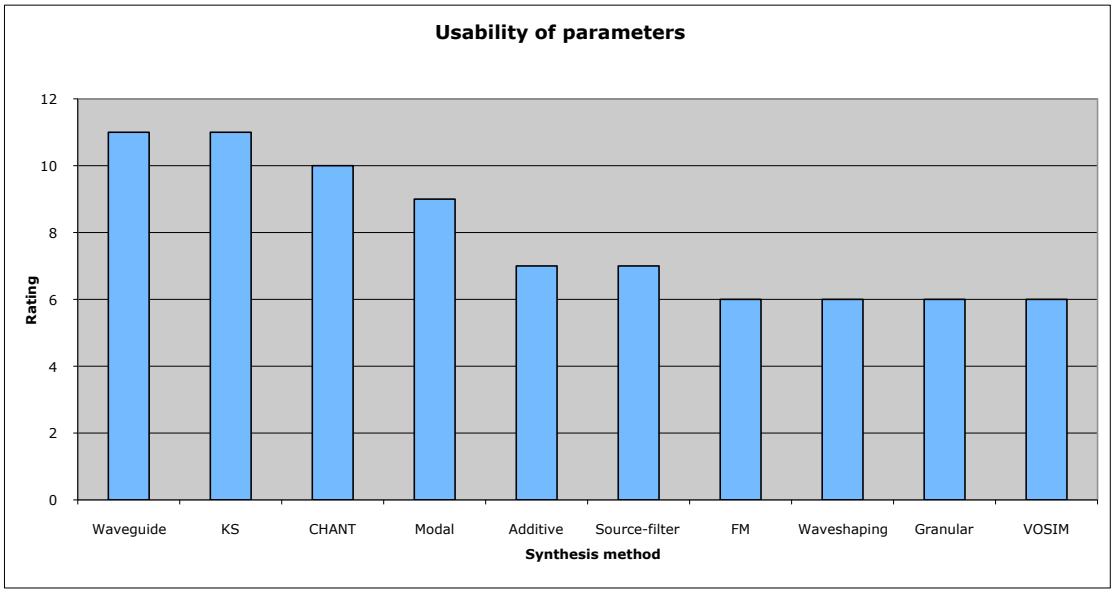


Figure 2.5: Synthesis parameter usability ratings (adapted from (Tolonen, Välimäki *et al.*, 1998)

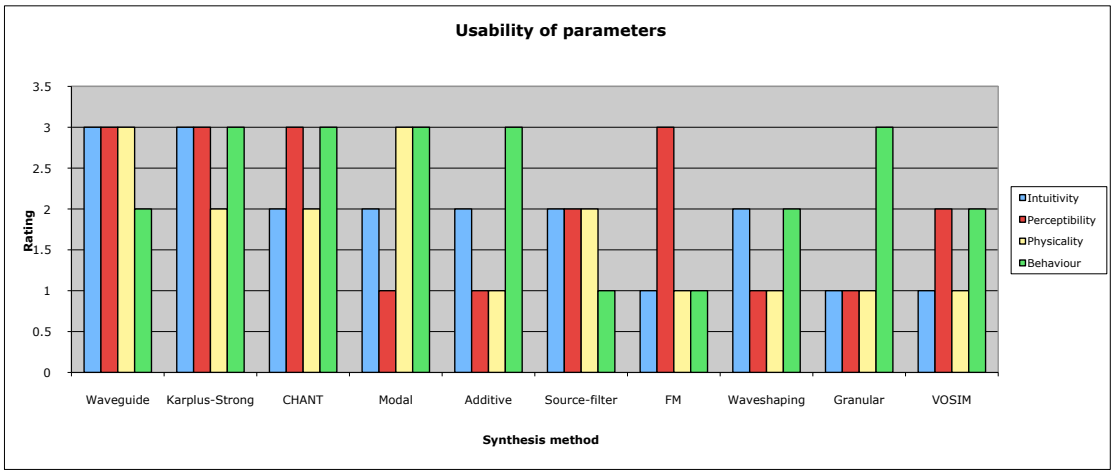


Figure 2.6: Synthesis parameter intuitivity, perceptibility, physicality and behaviour ratings (adapted from (Tolonen, Välimäki *et al.*, 1998).

In the view of Tolonen *et al* (and as shown in figures 2.5 and 2.6) waveguide, Karplus-Strong and CHANT synthesis score well in this survey. Waveguide synthesis, being founded on computer models of acoustic mechanisms, not surprisingly performs well on *physicality*, but precisely because it emulates the non-linear features of those mechanisms, the parameters are less 'well-behaved'. At the other end of the scale, the parameters of FM synthesis are rated *poor* on all measures of usability except *perceptibility* (a change in any parameter is certainly audible). Granular synthesis is also rated low except in

behaviour - changes in the sound are proportional to the degree of change in parameter values.

Figure 2.7 shows the ratings for the quality and diversity of produced sounds; figure 2.8 breaks this down into the subcategories of *robustness*, *generality* and the availability of *analysis methods*.

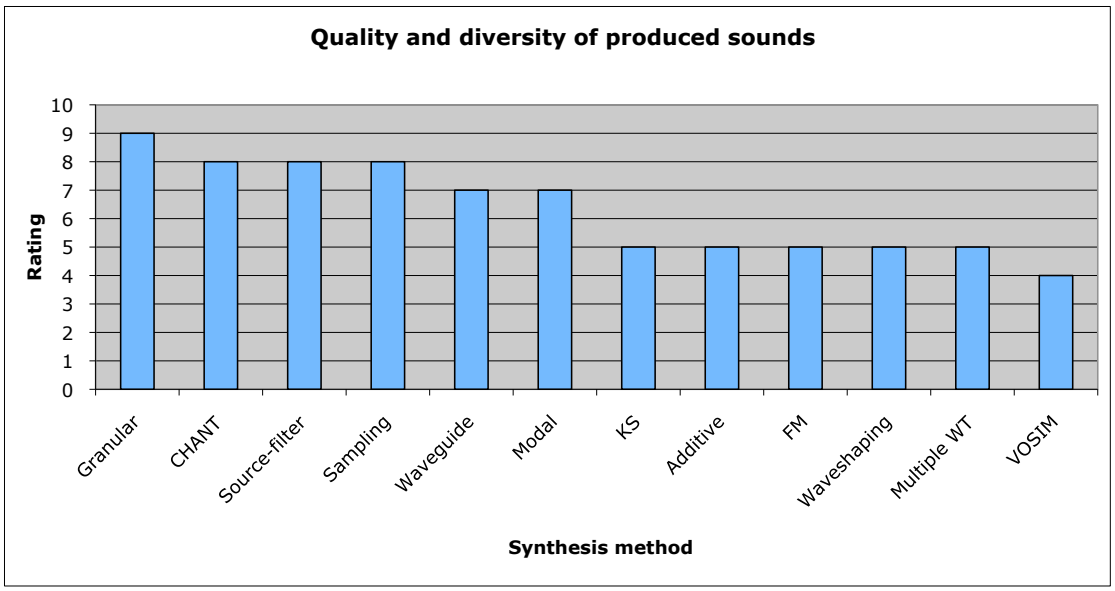


Figure 2.7: Ratings of synthesis method sound quality (adapted from (Tolonen, Välimäki *et al.*, 1998)

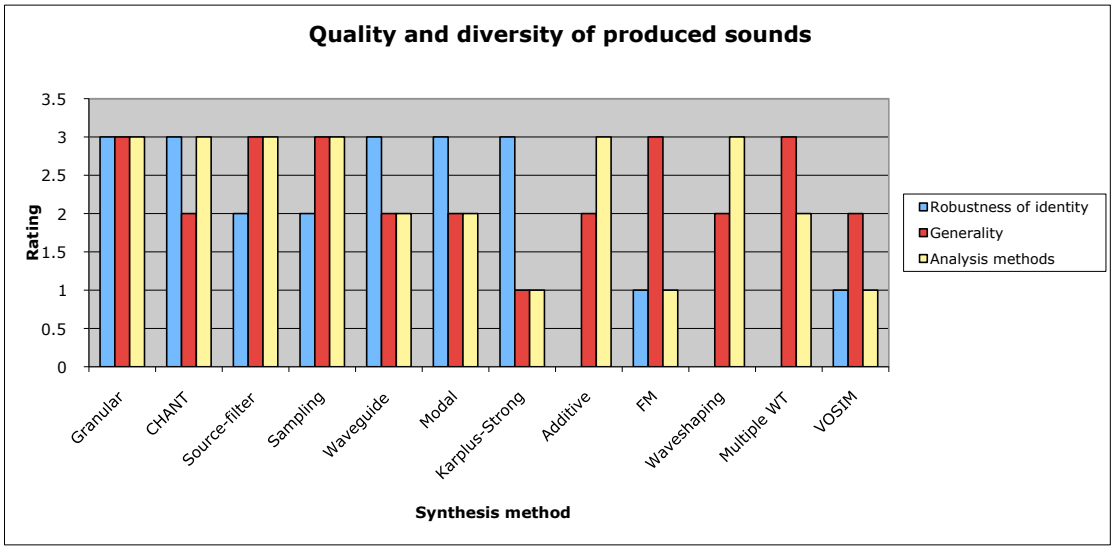


Figure 2.8: Ratings of synthesis method robustness, generality and analysis methods (adapted from Tolonen, Välimäki *et al* (1998)).

Here it is noticeable that the ranking is very different; there is no obvious correspondence, either positive or negative, between parameter usability and the quality and range of sounds available. Some methods score highly in all three subcategories (granular, CHANT and source- filter); others, such as FM, provide good 'generality', while performing poorly on 'robustness'. What emerges, however, from the survey is that all of these methods, to a greater or lesser extent, require users to have a clear understanding of the particular synthesis engine and its parameters, and to be able to convert an imagined sound into the correct parameter values. While 'expert' users may be able to do this, it is likely to be more difficult for those who are 'naive'. Some methods are more helpful than others – modal synthesis, for example, presents the user with a consistent metaphor which maps to real world phenomena, whereas methods such as FM and granular synthesis are less penetrable in this respect. Seen from an HCI perspective, there is a considerable gulf between the *task languages* familiar to a musician and the *core languages* inherent in synthesis methods. The following section explores this point in greater detail.

2.6. Synthesis hardware and software – an HCI perspective

2.6.1. Background

As noted at the beginning of this chapter, audio hardware and software for timbre shaping has received relatively little attention in the HCI research community. A 1995 analysis conducted on the working methods of composers working with Computer Music Systems (CMS) identified various typical tasks, and concluded that CMS designers must allow both for wide variations in composers' knowledge and skill and for wide individual variation in the types of composer they are designing for (Polfreman and Sapsford-Francis, 1995). Recommendations included: providing more than one level of interaction; hiding

unwanted levels of complexity; and employing knowledge based systems (KBS) to manage details that a user does not wish to specify directly. A previous critique of synthesizer user interface design (Ruffner and Coker, 1990) focused on the control surfaces of four contemporary instruments, and commented on the degree to which they conformed to design principles identified by Williges *et al* (1987). It was concluded that the demands placed on the user by the interfaces meant that they were far from ideal for the purpose: noting that, in general, 'user interface principles have been, at best haphazardly applied'. The authors also suggested issues that should drive future research in this area. Another more recent and related study (Fernandes and Holmes, 2002) has applied a heuristic evaluation to an electric guitar pre-amplifier interface.

The remainder of this chapter is devoted to the analysis and evaluation of a number of representative hardware and software synthesizer interfaces. We begin, however, by considering a general models of user-system interaction proposed by Norman (1988) and Dix, Finlay *et al (*1998*)*, and discusses the extent to which they can be usefully applied to the synthesizer interface.

2.6.2. The interaction framework

Norman's two-phase model of user-system interaction (1988) in computer systems is a convenient framework for considering the usability of the parameters of synthesis. This model consists of an *execution* phase, in which the user formulates a plan of action in pursuit of a goal and executes it, and an *evaluation* phase, in which the user compares the current state of the system with the desired goal state. This can be further divided into seven steps as follows:

Execution	Establishing the goal
	Forming the intention
	Specifying the action sequence
	Executing the action
Evaluation	Perceiving the system state
	Interpreting the system state
	Evaluating the system state with respect to the goals and intentions

Figure 2.9: The model of interaction (adapted from Norman (1988)).

The extent to which this interaction has been successful in a given computer system is determined by what Norman calls the *gulf of execution* – the difference between the actions as formulated by the user and the actions which are available on and permissible by the system – and the *gulf of evaluation* – the difference between the current state of the system and the expectation of the user. Where either of these distances is too large, the effectiveness of the interaction is likely to be poor.

An extension of the interaction model is one which represents the user (U), the system (S), the input (I) and the output (O), together with their associated languages (Dix, Finlay *et al.*, 1998), as shown in figure 2.10.

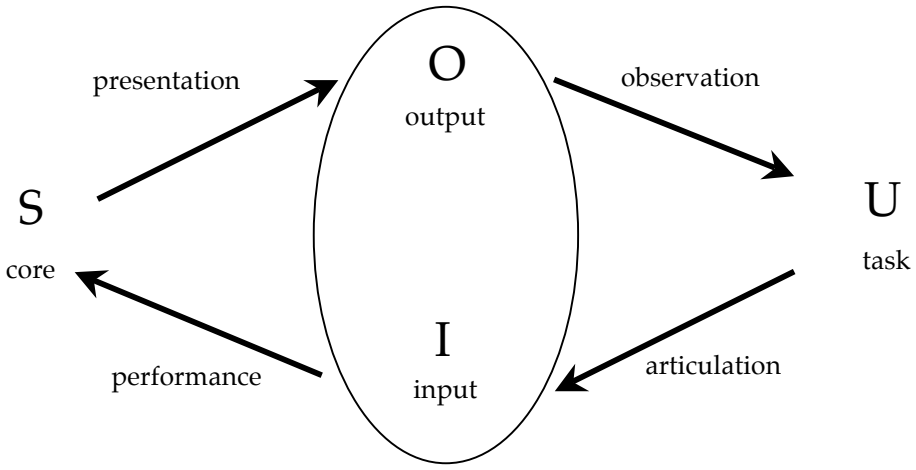


Figure 2.10: The interaction framework (from Dix *et al* (1998)).

In this model, the user formulates the goal using the terminology of the particular application domain – graphics, accounting etc; this is the *task language*. This goal is then *articulated* in terms of the actions available and permissible at the input. The input language is translated into the *core language* of the system (the *performance* phase), and the instruction executed. The evaluation phase then begins; the updated state of the system is *presented* to the user at the output, where the user *observes* and evaluates the results.

To what extent can this model be applied to the synthesis methods described above? For a musician, the *task language* might consist of adjectives and adverbs describing sounds - *shrill*, *spacious*, *dark*, *grainy* etc - or actions which would produce sounds – *plucked, struck, bowed* etc. The vocabulary of the *core language*, by contrast, refers to objective and measurable quantities associated with sound, such as spectral distribution and density, and their evolution over time. Input expressions in the *input language* may be different from either. As we have seen in the case of modal synthesis, for example, these will be couched in terms of density, stiffness, length etc, whereas grain size and envelope will need to be specified in granular synthesis. The extent to which a user is able correctly to formulate input expressions is clearly very much dependent on his or her level of expertise, which may vary between different synthesis methods. While the parameters of granulation or waveshaping, for example, may be abstruse to many users, they may very well be familiar with the terminology associated with subtractive (source-filter) synthesis. As previously noted, one striking aspect of the oscillator/filter/amplifier synthesis model associated with subtractive synthesis is the fact that it has survived the arrival of many other synthesis methods, and that it has in many respects become a *lingua franca* for audio synthesis. (In the user study reported later on in this chapter, a number of users were clearly confused by

the apparent absence of these modules in an interface which simply named them differently).

In general, however, the mapping of the *task language* familiar to musicians to the low-level *core language* which describes sound at an atomic level presents the 'naïve' user with a number of problems. Those synthesis methods rated by Tolonen *et al* as being highly usable (waveguide, Karplus-Strong and CHANT synthesis) nevertheless demand a high level of expertise for their effective use. The parameters of modal synthesis, already discussed above, provide a useful real world metaphor; however, as has been pointed out by Wessel, Risset and others, cited in (Chadabe, 1997), as far as sound generation is concerned, this particular technique restricts the musician precisely because of the metaphor. An imagined sound for which the composer can find no physical analogue is clearly one that cannot be easily realised using this synthesis method.

The above discussion of sound synthesis methods has drawn on the framework proposed by Dix *et al*, and has focussed on the usability of their various parameters (the *articulation* and *performance* phases), and has not considered the means by which the user is able to assess the effect of parameter value change (the *presentation* and *observation* phases). In order to do this, we turn now to consider some representative sound synthesis implementations.

2.6.3. Synthesis implementations

Over the past fifteen years, control surface designs of commercially available synthesizers have to some degree converged, to the extent that we can consider the instrument to have acquired a generic interface (Pressing, 1992). However, one cannot assume that similar looking buttons will perform the same function;

conversely, a given function could be performed by a number of different controls. (Thimbleby's (2001) example of the design of electronic calculators is of relevance here. He notes that the hand held calculator is a 'mature technology', with well defined requirements, but goes on to describe two models of calculator which look superficially very similar, but whose controls often do different things.) As noted in the opening chapter, cost restraints imposed on hardware synthesizer manufacturers, together with the limited space available on the control surface could account for this to some extent, but does not explain why such limitations have been exported to the software equivalents.

Pressing (1992) describes the controls of the synthesizer user interface as falling into two broad categories: those which govern 'real time' synthesis, and those which provide access to the parameters governing 'fixed synthesis'. Real time synthesis controllers, such as pitch wheels, foot pedals and the keyboard, allow instant and dynamic modification of single scalar aspects of existing sounds: frequency, filter frequency, amplitude etc. These controllers are designed and positioned on the control surface to meet the requirements of real-time performance, and it is relatively easy for users to understand their use: the effect that a controller has on the sound is instantly audible.

The part of the interface devoted to 'fixed synthesis' is the focus of the remainder of this chapter. The 'fixed synthesis' component of the interface allows the design and programming of sound objects. Its informed use typically requires an in-depth understanding of the internal architecture of the instrument, and the methods used to represent and to generate sound. Thus, under most current systems, the user is obliged to express directives for sound specification in an *input language*, rather than in language derived from the user domain.

2.6.3.1. Fixed synthesis user interface designs

Since the development of the early synthesizers of the 1960s, based on analogue electronics, three distinct interface architectures have emerged for fixed synthesis. In approximate order of the complexity of associated user interface issues, (though not necessarily their complexity from other perspectives) they are as follows.

- Parameter selection in a fixed architecture;
- Architecture specification and configuration;
- Direct specification of physical characteristics of sound

For purposes of exposition, and reflecting historical trends, it is useful to begin with the second of these approaches first: *architecture specification and configuration*, also known as *user specified architecture*. This approach to specifying timbre has its origin in the interfaces of early synthesizers, such as the Arp, Moog and EMS. In such early synthesizers, a given sound was defined in terms of the configuration of electronic modules required to generate it. The hardware interface offered total control over the choice, interconnection, and settings of these modules via physical plugboards or patchbays. Modern versions of this idea use GUI based interfaces to accomplish similar ends.

The approach appearing first in the list above (*parameter selection*, also known as f*ixed architecture*) came next historically. This approach effectively froze, or pre-patched, selected configurations of modules and simply allowed the user to vary the values of parameters controlling these modules. Different synthesizers may use quite different sound synthesis modules from each other, but the principle remains the same. Thus, *fixed architectures* present to the user an internal model of

sound which is essentially a tree or graph structured assemblage of parameters. For the user, the task of defining a sound is one of traversing this structure, specifying parameters e.g., by a 'form filling' process. The earlier mentioned user specified architectures, by contrast, are essentially fluid and non-hierarchical. We will revisit both types below.

Finally, the third category of user interface for timbre control in synthesizers is *direct specification*. First widely introduced commercially in early Fairlight synthesizers, it allows the user, in principle, to specify sound directly by, for example, drawing or modifying a waveform on the screen.

The next three subsections 2.6.3.1.1 to 2.6.3.1.3 will consider each of the three categories in more detail, describing modern interfaces from each category. We will draw on a series of user tests comparing the categories (Seago, 2004).

2.6.3.1.1. Parameter selection in a fixed architecture

As noted above, the *fixed synthesis* control surfaces of more recent hardware-based synthesizers (recall that *fixed synthesis* does not mean *fixed architecture*) have standardised in recent years. Typically, there are selection controls for preformatted sounds (known as 'programs' or 'patches'), programming controls (to change program parameters) and mode selection controls (play, edit, etc). Limitations on control surface space mean that controls may be multi-functional: their usage at any given time will be determined by the mode currently selected.

The model of sound generation used in interfaces of this category has a static and hierarchical structure, whose constituent parts are typically (but not

always) the parameters of subtractive synthesis, with settings defining waveforms, envelopes, filter cut-off frequencies, etc. The task of defining or editing a sound involves the traversal of this structure, incrementally modifying the sound by selecting and changing individual parameters. Examples of such an interface are those of the Yamaha DX7, offering FM synthesis, and SY35 (wavetable and subtractive synthesis). The LCD indicates no more than one parameter at a time, providing no overall visibility of the system state. However, since all parameters have default values, instant feedback is available simply by listening to the current sound; the user is able to assess the effect of the changes made; actions are at all times reversible, and errors or 'illegal actions' are impossible. Parameters are selected, and modifications effected, in the same way throughout the structure.

2.6.3.1.2. Architecture specification and configuration

In this architecture, sound is viewed as the output of a network of functional components - oscillators, filters, and amplifiers. The structure of this network is fluid, and can become quite complex. The output of any element may be processed by one or more other elements. However, even greater fluidity comes from the fact that the parameters of each element, frequency, envelope and cut-off frequency, etc, can be dynamically controlled by the output of any other element. As already noted, early subtractive synthesizers were in this category; the basic components were linked by physical patch cords, and the signal path was visible and immediately modifiable.

In hardware synthesizers, the range of sound that can be produced is limited by the number of hardware modules available. Software versions, however, in important respects, have no such restrictions. *Reaktor 5* (2009) is a good example of a synthesizer that emulates and mimics in software a modular

subtractive synthesizer (it also offers fixed architecture configurations as well). A more detailed examination of this particular configuration example will be made later in the chapter.

2.6.3.1.3. Direct specification

All the user interfaces examined in the previous two sections are predicated on a model of sound as an assemblage of components which generate or modify sound. This assemblage, having been designed, is the engine which generates the required sound. The following section deals with *direct specification* interfaces that allow the desired sound to be specified more directly.

The notion of *direct specification* has much in common with *direct manipulation* or *DM* (Shneiderman, 1983; Shneiderman, 1997). This interaction technique, which is now the basis for all modern graphical user interfaces, has a number of characteristic features:

- Visibility of the object of interest (e.g., documents, folders etc)
- Incremental action at the interface with rapid feedback on all actions
- Reversibility of all actions, so that users are encouraged to explore without severe penalties
- Syntactic correctness of all actions, so that every user action is a legal operation
- Replacement of complex command languages with actions to manipulate directly the visible objects

One of the most important aspects of a DM interface is that it is less easy to make a clear distinction between input and output objects. The *output expression* of the system is capable of being used to formulate a subsequent *input expression.*

At first sight, it would seem that direct manipulation interfaces lend themselves well to effective user specification of sound. Time domain plots could serve both as output and input objects, with tools provided to 'draw' and 'edit' the desired waveform. (Such an interaction style would be well suited to wavetable synthesis methods.) Similarly, for additive synthesis, the relative amplitudes of the harmonics of the frequency spectrum could be adjusted to produce the desired sound (such an interface is reviewed both in this chapter and in chapter four). However, a user interface for 'designing' sounds in any detail in this way is hampered by the lack of any human understandable mapping between the subjective and perceptual characteristics of the sound in any detail and its visual representation on screen.

The problems of audio-visual mapping for sound visualisation are reviewed in Giannakis (2006). Visual representation of sonic information is typically in either the time domain (a plot of amplitude envelope with respect to time,) or the frequency domain (a plot of the relative amplitudes of the frequency components of a waveform. The interpretation of time domains plots such as that shown in figure 2.11 is, to a certain extent, intuitively clear; this is a sound made up of sonic fragments, of varying degrees of loudness, punctuated by silence (this is a recording of normal speech).

Figure 2.11: Time domain representation of sound.

However, and crucially, reconstructing the sound from this visual representation would be, for all intents and purposes, impossible. In order to accomplish this, the user would need to zoom in on this display to reveal its instantaneous waveform. However, it is equally difficult to make an intuitive association between the waveform and the sound it generates; the information is simply at too low a level of abstraction. In addition, the mapping of a sound to its waveform is not unique; two waveforms with similar spectral envelopes but with differing phase spectra will sound identical, but look different (Roads, 1996). Frequency domain representation takes the form of a frequency spectrum plot derived from Fourier analysis of a waveform. Again, the *core language* information which it provides on the frequency content of a waveform is too low-level to be useful as a general means of manipulating sonic attributes. (At best, the user would need to apply heuristics such as 'a brighter sound has more energy in the upper frequencies.) A third form of representation (also in the frequency domain) is the sonogram, a frequency against time graph that presents the time-varying spectrum of a sound or sequence of sounds. Typically, colour is used to represent the amplitude of the spectral components. The advantage of this representation is the visibility of change with respect to time – however, the issues of specification apply with equal force here.

In practice, no user is able to specify finely the waveform of imagined sounds in general, either in the time or frequency domains. In other words, there is no *semantic directness* (Hutchins, Hollan *et al.*, 1986) for the purpose of specifying any but the most crudely characterized sounds. The gap between core language and task language is just as wide as in the first two categories. It is for this reason that the term *direct specification* will be used here, in preference to *direct manipulation*.

2.6.3.2. Task analysis and heuristic evaluation

In order to make the above discussion more concrete, this section reports on a *task analysis* and *heuristic evaluation* carried out on a small but broadly representative range of synthesizer user interfaces. The synthesizers surveyed here are discussed under the three interface categories noted above.

Task analysis is simply the decomposition of a task into its constituent sub-tasks, each of which, in turn, may consist of a number of sub-sub-tasks (Dix, Finlay *et al.*, 1998). The output of the process is a hierarchical listing of all the tasks required in the order that they are performed. The task in this case is the creation of a simple 'sound object'; a sound whose time domain waveform is a sawtooth, whose frequency is 440 Hz, and whose overall amplitude envelope describes a long and smooth decay; a task analysis was performed for each of three synthesizers – the Yamaha *SY35*, a hardware synthesizer whose user interface can be broadly described as being 'fixed architecture'; *Reaktor*, a 'user specified architecture' software synthesizer ; and *Metasynth*, also a software synth, whose interface is one which can be characterised as 'direct specification'.

Each step in the task analysis was considered using a *heuristic evaluation* technique. This is a structured method for evaluating a user interface design against a number of design principles and criteria (Nielsen, 1994). These are listed in figure 2.12.

Visibility of system status	At any time during the interaction, the user should be able to assess the system state from its interface representation.
Match between system and real world	The interaction should, as far as possible, be expressed in terms of the user domain.
User control and freedom	Actions should be reversible; the user should be able to change his/her mind.
Consistency and standards	User/system communication should always be done in the same way
Error prevention	The possibility of error should be designed out ; where this is not possible, the system should provide good error messages, and enable the user to recognise, diagnose and recover from errors
Recognition rather than recall	Objects and actions should be visible; the user should not have to remember which actions are available at any given time
Flexibility and efficiency of use	There should be shortcuts for expert user
Aesthetic and minimalist design	The interface should present only that information which is relevant at that time, and should not include irrelevant information
Help users recognize diagnose and recover from errors	Error message should be expressed in plain language

Figure 2.12: Heuristics for interface evaluation - (adapted from Nielsen, 1994)

In order to highlight some of the features of, and differences between various interaction styles, an adapted version has been used (normally, such an evaluation would be carried out by several independent evaluators).

2.6.3.2.1. Comments on methodology

The evaluation criteria proposed by Nielsen highlight a number of general issues to do with the specification of sound, which will be considered here.

Software sound synthesis invites comparison between the types of interactions it affords, and those of other application domains, such as text processing, graphics and animation. However, the user interface of the typical commercial synthesizer, in both software and hardware, more closely resembles those interfaces which facilitate industrial control processes, which act as intermediary between the operator and the real world (Dix, Finlay *et al.*, 1998), and

in which two levels of feedback are provided; firstly, from the instrumentation - meters, gauges etc; and secondly, from the equipment or processes being controlled (see figure 2.13).

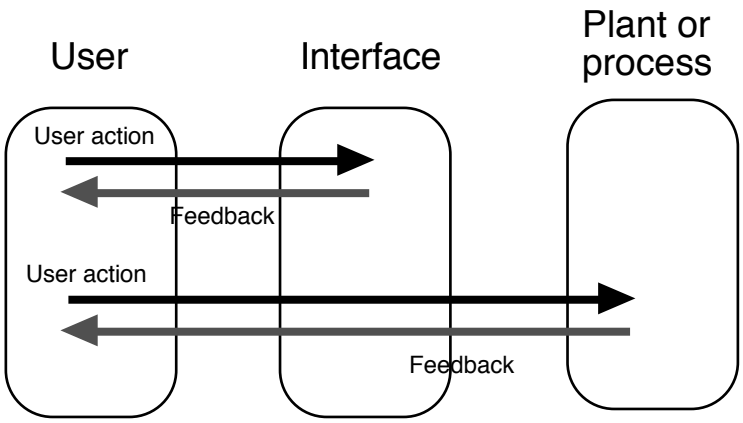


Figure 2.13: Indirect manipulation; adapted from Dix *et al* (1998).

A useful analogy is that of the flight control deck, where the pilot is provided with specific information on altitude, pitch, yaw, and speed from the instrumentation, but also is more directly afforded 'real world' information by the behaviour of the aeroplane itself. Similarly, the user of a typical synthesizer is able to determine the current 'state' of the sound being edited by the parameter values displayed in the LCD, but can also audition the effect of the editing process by pressing a key.

This is not the case with, for example, a desktop operating system, where the mouse dragging of a document icon from one folder to another is not accompanied by any perceptible 'real world' feedback - because the user is engaging with a "representation or model of reality" (Shneiderman, 1997), he/she takes it on trust that the action has actually happened. Similarly, the results of actions taken by the user of a WYSIWYG graphics or word processing program are visible on the screen - no confirmatory 'real world' feedback is provided until the document is printed. From the user's point of view, the interface and the objects of interest are one and the same.

Such a 'fusing' of the object of interest and its iconic representation is more problematic with sound synthesis systems, most obviously because the object of interest is non-visual. While an iconic representation can readily be found both for objects such as documents, drawing tools and geometric shapes, and for actions such as 'copy', 'delete', 'draw' etc, sound does not lend itself to any form of consistent visual description which can be said to make 'psychological sense' (Karkoschka, 1966) – that is to say, where its subjective and timbral, rather than purely acoustical and measurable characteristics are apparent. The reasons for this will be examined in much greater detail in the next chapter; however, the separation of the object of interest and its interface representation has implications for two of the criteria on which a heuristic evaluation of this type is conducted.

The first of these is *visibility of system status*; the stipulation that the user interface should at all times provide the user with feedback and information on the current state of the system. In a synthesizer, such feedback is likely to be on two levels – the 'interface' level – current parameter values for example - and the 'plant or process' level – the sound itself. This being the case, evaluation of the three synthesizers under this criterion will reflect these two levels.

The second heuristic evaluation criterion, *match between system and real world*, is about the extent to which input and output expressions are couched in the *task language* rather than in the core language of the system. Again, this is less clear cut in sound synthesis than is the case in other application domains, not least because it is not obviously clear what the task language is. If its lexicon is the properties and attributes of 'real world' acoustical components and mechanisms, then clearly modal synthesis excels in this regard. If the language of voltage controlled subtractive synthesis has become a *lingua franca* for synthesis (as previously

suggested), then a successful interface clearly will be one that presents representations of voltage-controlled oscillators , low frequency oscillators and voltage-controlled filters to the user. If, on the other hand, the task language draws on (for example) a musician's vocabulary – *bright*, *rough*, *sharp* etc - (as was mentioned in the discussion on the interaction framework – section 2.6.2), virtually all synthesis methods and implementations perform poorly.

It should be emphasised that the intention here is to focus on and highlight those HCI features that are characteristic of, and intrinsic to these three architectures, rather than to use all of these criteria to analyse each implementation exhaustively to identify minor usability issues, which could in principle be easily corrected. It should also be noted that many synthesizers do not fit solely into one or other of the three categories described above, and make use of a variety of techniques to achieve the same end; indeed, one of the features of a good interface is precisely this kind of flexibility. Lastly, the methodology presented in the following section assumes that the use of synthesizer timbral editing controls is always target-oriented - that is to say, the user's focus is on either creating an imagined sound or recreating an already existing one. However, this is not always the case: as has been noted in a recent study, 'exploratory and improvisational uses also exist' (McDermott, 2008). These two modes of navigating the synthesis space are not necessarily conducted with a particular target sound in mind.

This poses a problem; as McDermott observes, a system designed to facilitate exploratory and improvisational usage does not, by its very nature, allow the definition of performance metrics, and is therefore not easily testable. On the other hand, the (re)creation of an imagined or previously heard sound is a realistic task; and performance metrics and indicators can, by contrast, be

readily specified for this mode of operation. For this reason, McDermott's system (discussed elsewhere in this thesis) was based on an assumed target-driven mode of operation. Both the task analysis and heuristic evaluation presented here, and the design and testing of the systems described in chapters six and seven are similarly founded on this assumption; the limitations of such an approach, however, are noted.

2.6.3.2.2. Comments on the task

All three of the synthesizers examined here are very different in their capabilities, performance modes, and methods of sound generation, and in order to compare 'like with like', the chosen task is deliberately limited in scope.

It is also, to some extent, contrived. In a working situation, a user would be more likely to take an existing sound from the library available, and edit it, rather than generate a sound *ab initio*. However, given that the libraries available will differ (and will not exist in some cases), the evaluation necessarily needs to analyse the process from the beginning. Within these limitations, the task is nevertheless realistic, and one which might be undertaken by a user.

Five distinct phases in the generation process can be discerned; *initialisation*, *waveform selection*, *pitch selection*, *envelope selection*, and *save*. The first and last of these really only apply to synthesizers which are programmable; however, the second, third and fourth phases (which may occur in any order) are common to most, if not all architectures, and relate to timbre, pitch and loudness respectively. It is the first of these attributes of sound and the limitations of existing controllers for it that is the focus of this thesis, and the subject of discussion later in this chapter.

Each synthesizer is evaluated both in terms of the number of steps required to complete the task outlined above, and against the criteria listed in figure 2.12. The raw data from each task analysis is viewable in the *Task_analysis* folder on the CD (I. Task_analysis/introduction.html).

2.6.3.2.3. Yamaha SY35

The user interface of this synthesizer has a hierarchical structure as shown in figure 2.14.

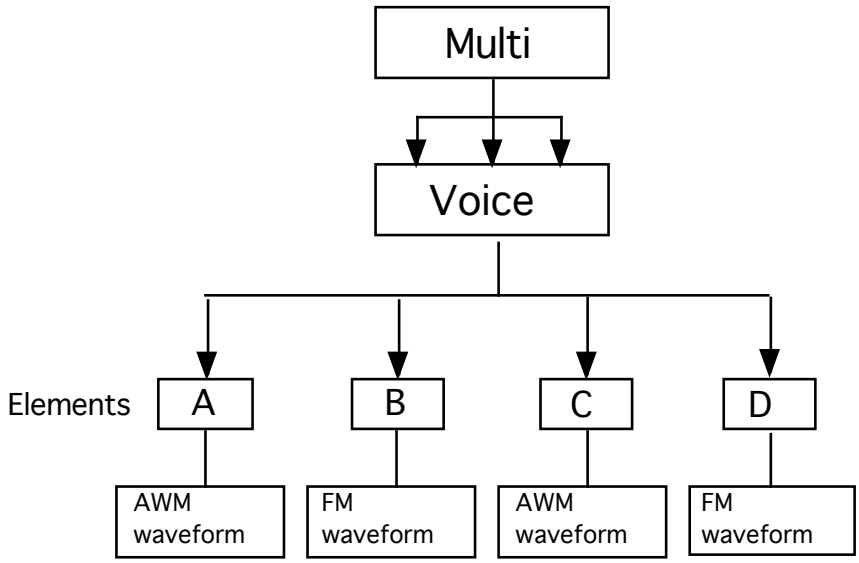


Figure 2.14: Hierarchical architecture of the Yamaha SY35.

A given instrument (or *multi*) consists of a number of voices, each of which comprises two or four waveforms (or *elements*). There are two synthesis engines, one based on FM, the other on Advanced Wave Memory (AWM) - essentially a wavetable method). Parameters can be specified/modified at all levels of the structure; pressing the Edit/Utility button provides access to these parameters.

The *initialisation* phase requires four separate actions in all. The selection of the waveform involves cycling through the available options, using the +1/YES

button, using an arrow button to locate the cursor on the correct field, and then setting the value of the selected field, using the +1/YES button again. The number of actions required is at least nine, and may be more; specifying the amplitude envelope involves even more actions. Overall, the process involves forty-seven steps.

The modification of a sound can only be done incrementally. The LCD indicates no more than one parameter at a time; this means that there is no overall visibility of the system state at 'interface' level. At 'plant or process' level, however, constant feedback is available; the user is able to assess the effect of the change that s/he has made. Actions are at all times reversible, and errors - 'illegal actions' - are impossible. Parameters are selected, and modifications effected in the same way throughout the structure.

It is interesting to consider this 'tree' structure negotiated by the user in the light of 'depth versus breadth' studies of menu structures in application software packages (Kiger, 1984; Landauer and Nachbar, 1985; Norman and Chin, 1988). Menu structures can be either 'deep', with a number of levels, or 'broad' with fewer levels, but there seems to be general agreement that users are better able to navigate 'broad' structures with no more (and preferably fewer) than four levels. The menu structure on the SY35 conforms to the 'broad' model.

2.6.3.2.4. Reaktor

Reaktor (Native Instruments, 2003) is a software synthesizer that, among other things emulates and mimics a modular subtractive synthesizer. The organising principle here is one in which individual software 'synthesizers', or *instruments* are grouped to form an *ensemble* (shown here in Fig 2.15).

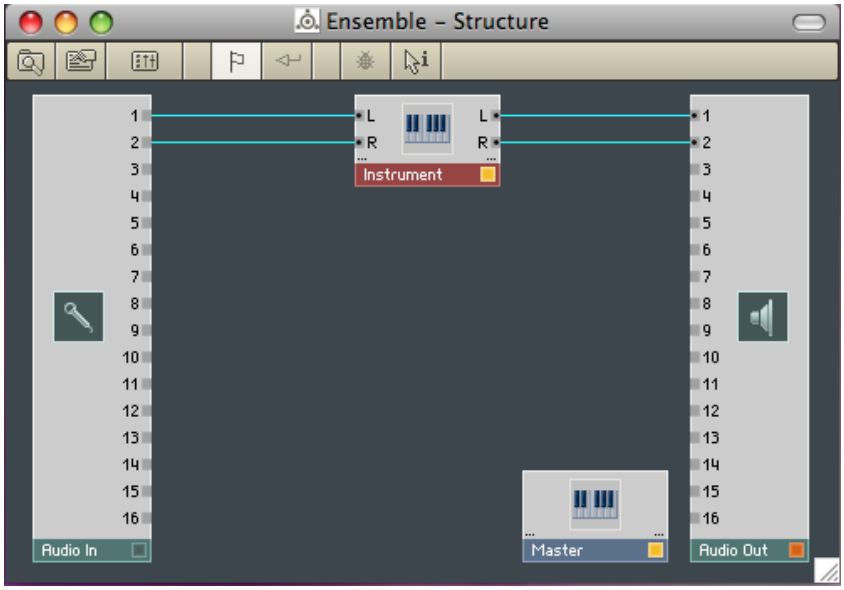


Figure 2.15: Reaktor – ensemble structure.

Each instrument is made up of a number of modules, some of which may be drawn from the subtractive/FM synthesis domain (envelope generators, oscillators, etc); others may themselves be instruments. Connections between components are made by mouse-dragging, and in this way, a complex and fluid structure may be generated; one which is also recursive, in that instruments may be defined as assemblages of other instruments (see Fig. 2.16)

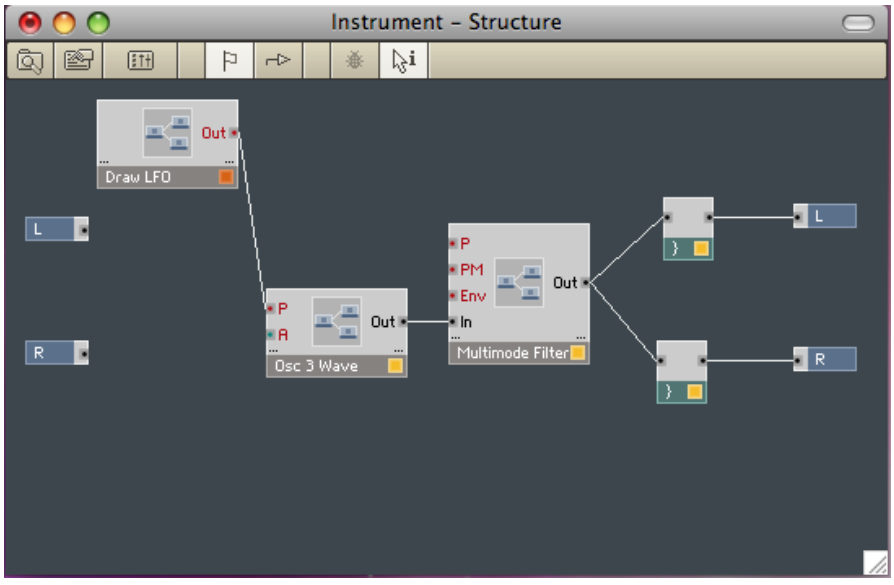


Figure 2.16: Reaktor – instrument structure.

Four actions are required to initialise an instrument, two to select the waveform, and two for the pitch. The process of specifying the envelope, however, is rather more complicated, and most of the interaction is devoted to this area.

Four actions are required to complete the voice initialisation phase, and two each to select the waveform and pitch. However, no less than fourteen actions are required to specify the required envelope, and overall, twenty-two discrete steps are required to complete the task.

Visually, the interface presents a clear and uncluttered view of the system. As in the hardware version, there is clear visibility of the system state at all times at 'interface' level. However, at 'plant or process' level, like the hardware version, the user is unable to aurally evaluate the success of his/her actions until a minimum number of connections have been made; up until this point, there will be no sound at all. Actions throughout are reversible, the interaction is consistent throughout, (a given action will produce the same result in different contexts), and the GUI makes 'illegal' actions impossible.

The interaction involved in building an instrument is one of direct manipulation; it is important to emphasise however, that the 'objects of interest' with which the user engages are not representations of the sound itself, but of the functional components required to create it; it therefore differs from the 'direct specification' interface of *Metasynth*, to which we turn.

2.6.3.2.5. Metasynth

Metasynth, produced by U and I Software, is a package for sound design and the creation of electronic music; for the purposes of discussion, we will focus

on the Image Synth, which provides facilities for the synthesis of single sounds, musical phrases or entire soundscapes. Like *Reaktor*, it makes use of a direct manipulation interface, but more specifically, it is a 'direct specification' interface in that the user is interacting with a visual representation of the sound itself rather than with a representation of the functional components needed to generate it.

The user interface takes the form of a pitch/time graph; each line represents a sound which may be a sound produced by a synthesized instrument (effectively making the display a kind of musical score). Alternatively, a line may represent a single partial (in which case the display is a sonogram). In either case, the amplitude of a component at any moment is indicated by its brightness. The user is provided with a palette of tools for editing this image; a new sound can be auditioned at any time by generating the sound depicted by the image. The frequency against time representation of the sound has the potential to make the perceptual mapping of the image and the attributes of the sound it represents rather more apparent than is the case with a purely time domain representation (however, note the discussion in the next section).

In principle, the initial part of the test task - the creation of a sawtooth waveform - should not pose any great problems. The harmonics of the sound can be drawn into the window, and the amplitudes of each adjusted by the shading tools. Achieving the goal in this way is quite laborious, however, and the program does offer other, more direct ways of achieving the same end. The Wavesynth window, shown in figure 2.17 and which is essentially a wavetable synthesis implementation, enables the user to specify and graphically edit a waveform from a palette of waveform archetypes. (It should be noted, however, that the waveform labelled 'sawtooth' in this window is actually a 'triangle' wave, and the user needs to make use of another waveform which more closely resembles a

'sawtooth'. This is an error of implementation, however, and does not have a bearing on the discussion presented here.)

Figure 2.17: Metasynth – the Wavesynth window.

Having selected the waveform, the user specifies the pitch by selecting the drawing tool, clicking on the vertical position of the Image Synth frequency/time display corresponding to 440 Hz. A line is drawn across the screen. The waveform can then be generated, and a time domain representation appears at the top of the screen, as shown in figure 2.18. The envelope specification is done by repeated clicking on the 'fade out' button, and the resultant sound file saved in the standard way.

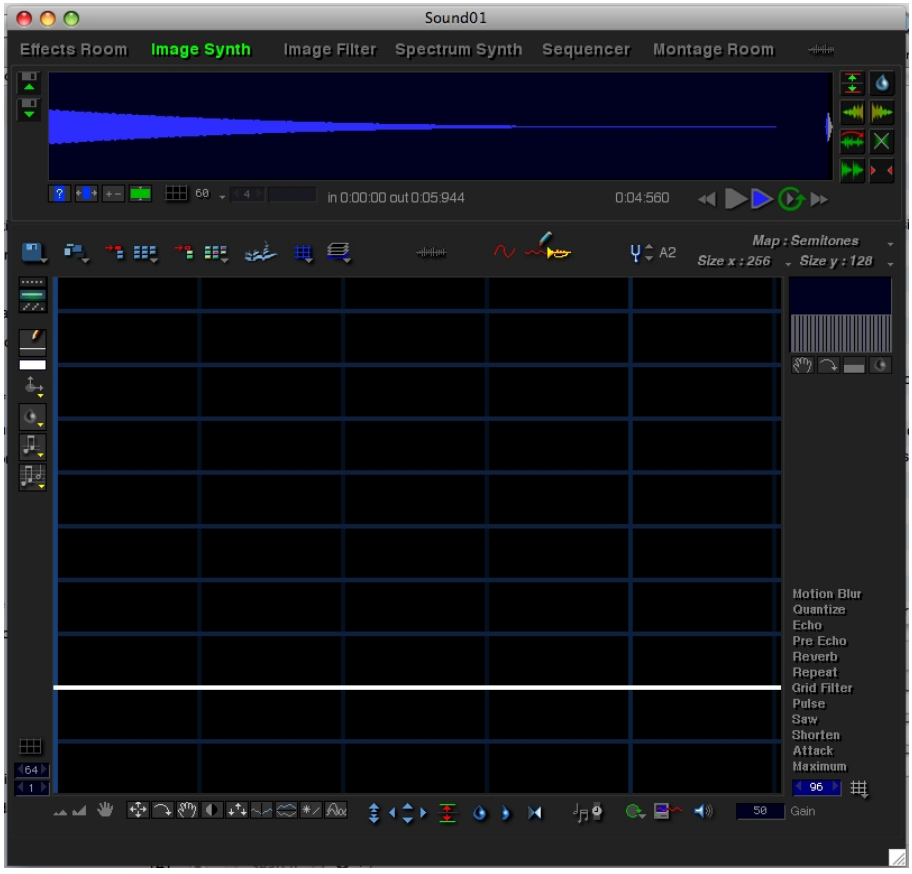


Figure 2.18: Metasynth – the ImageSynth window.

In general there is visibility of system status throughout – the change in the envelope is made visible in the time domain, the change in waveform can be both seen and heard in the Wavesynth. (In this particular application, however, the Image Synth display is ambiguous and inconsistent in what it displays, not necessarily providing the user with a visual depiction of the changing spectrum over time.) Insofar as the vocabulary of the task language is considered to be that of waveforms, filters and frequency spectra, there is a good match between the task language and the system itself.

Expressed purely in terms of the number of user actions required to accomplish this particular task (including the saving of the voice to a file), *Metasynth* ('direct specification' architecture) is the most straightforward to use, requiring eleven mouse clicks in all. The interface of the Yamaha SY35 (*fixed*

architecture) - emerges as the least well suited to the test task, requiring no less than forty-seven discrete steps.

2.7. Conclusions and discussion

A number of synthesis methods and implementations have been evaluated for usability in this chapter. Synthesis methods can be classified into four main categories - *abstract algorithms, spectral models*, *sampling* or *processed recordings* and *physical models* (Smith, 1991). In addition, we have reviewed and classified a number of synthesizer user interfaces, and a taxonomy of interaction styles emerges. The first is one in which a sound is presented as the assemblage of components required to generate it; in the second model, users navigate an already existing assemblage of procedures (fixed architectures), typically by form-filling and/or menu selection. The third, *direct specification*, has many of the features of direct manipulation; however, the aspect of it which distinguishes it from the first is that the user engages with a visual representation of the sound, or some part of it. It is instructive to map these two taxonomies on to the other.

Interface architecture	Synthesis method	Synthesis type
Fixed architecture	**FM**	**Abstract**
	Waveshaping	
	Karplus-Strong	
	Granular	**Sampling/processed recording**
Direct specification	**Sampling**	
	Multiple wavetable	
	Additive	**Spectral**
Architecture specification	**Formant synthesis**	
	Subtractive	
	Modal synthesis	**Physical models**
	Digital waveguide	

Figure 2.19: Classifications of synthesis methods.

This chart is clearly indicative rather than definitive, and other classifications are possible; as was noted earlier in this chapter, most, if not all of these synthesis methods could in practice be implemented in more than one of these architectures.

The taxonomy groups the synthesis methods according to the interface architecture to which they would, nevertheless, seem to be best suited. FM synthesis, for example, could not be easily realised in a *direct specification* (as distinct from a direct manipulation) architecture, whereas additive and wavetable synthesis invite such an implementation.

What is noticeable, however, is that the synthesis methods for which the task languages are most decoupled from 'real world' or 'musical' associations and terminologies – the *abstract* methods – tend to be those that are realised using *fixed architecture*s where the mode of interaction is one of extended menu navigation and form filling. By contrast, sampling/processed recording and spectral synthesis approaches, whose task languages map more readily to measurable properties of sound can be achieved using *direct specification* methods, which the above heuristic evaluation suggests is a more useful and intuitive means of specifying sound.

The potential advantages of direct specification methods over fixed architectures were further evidenced in a series of user tests in which a number of subjects, undergraduate students of music technology at London Metropolitan University, were asked to perform three tasks on each of two commercially available hardware synthesizers (Seago, 2004). The interface of one of the synthesizers, a Roland XP50, was of the fixed architecture type; that of the other, a Korg Trinity, incorporated some elements of direct specification. The tasks were as follows:

- The selection of a particular sound, or 'patch' from the available library of preset sounds. In this study, the sound was that of a piano.
- The modification of the volume 'envelope' of that sound, such that, instead

of beginning suddenly and percussively, the sound started from inaudibility and increased in volume to a maximum over the period of about a second.

- The modification of the 'tone' of the sound, making it sound 'brighter'.

At each stage of the interaction, subjects were asked to describe aloud what were thinking - what they were trying to do, what questions and problems presented themselves, and what they were inferring from the current state of the interface. In many cases, subjects required prompts in order to complete the task successfully. In general, subjects found the tasks, particularly the third one, easier to accomplish on the Korg Trinity (the direct specification type), and when asked, expressed a unanimous preference for this synthesizer. Why is it, then, that *direct specification* methods are not universally used for sound synthesis?

Firstly, the sound object created in the heuristic evaluation task was very simple; a more complex sound would be less easy to create using the techniques characteristic of the direct specification approach. To create a time variant sound *ab initio* using *Metasynth's* Image Synth window shape would be quite a protracted process, requiring the 'drawing' of individual partials and specification of their varying frequencies and amplitudes with respect to time. More importantly, the specification of a sound requires the informed use of a task language whose vocabulary is based on this level of specification. The user has to be able to express the goal in terms of partials, spectrum, envelope etc; even for an informed user, this may be difficult.

Not all aspects of sound are difficult to specify, of course. Pitch (if the sound is pitched) and loudness can be mapped more or less directly to fundamental frequency and amplitude respectively, and appropriate controllers easily devised. Similarly, the overall amplitude envelope of a sound – whether it starts and finishes suddenly or gradually - can be broadly captured using the ADSR

controllers of voltage controlled synthesis. Only the colour of the sound – its timbre - remains elusive and multi-dimensional, much less easily captured and represented in a way that allows intuitive modification.

Furthermore (as discussed in section 2.6.2), the task language for the musician is more likely to be descriptive rather than prescriptive. The left hand box of figure 2.20 lists examples of terms which might be used by a musician to describe a given sound.

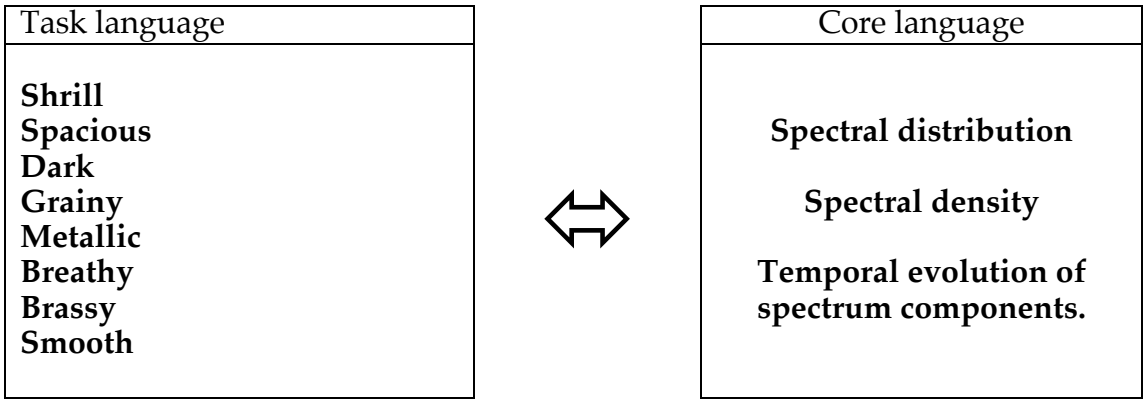


Figure 2.20: Task and core languages in synthesis.

Such terms are used to describe those attributes of sound - timbre, texture and articulation - which cannot be entirely captured by the conventions of the Western musical score and are often chosen for their perceived extra-musical analogies with other domains - colour and texture, for example - or for emotional associations. The right hand box, by contrast, describes objective and measurable quantities and attributes associated with sound. The mapping of one set of descriptors to the other is a problem for psychoacoustics and computer music researchers alike, particularly (as will be shown in the following chapter) as there is frequently no common understanding of these subjective terms.

The following chapter examines timbre from both a musical and psychoacoustic perspective, and reviews key research in the area.

Chapter 3 - Timbre and timbre space

3.1. Introduction

Because this thesis is primarily concerned with the tools available to musicians and composers for timbral shaping, this chapter will examine timbre as a musical resource, contrasting approaches to its classification and taxonomisation and finally research that has been conducted over the past forty years into its psychoacoustical basis.

The study of timbre – the 'colour' or 'quality' of sound - spans a number of disciplines - music, psychoacoustics, acoustics, linguistics, cognitive psychology, neurology and evolutionary psychology, the first two of which are the most pertinent to this study. While timbre is a compositional resource, increasingly foregrounded in Western music from the nineteenth century onwards (particularly in electroacoustic music post-1945), it is, at the same time, a perceptual and psychoacoustical phenomenon, and the determination of its salient characteristics an empirical problem for researchers in psychoacoustics and the psychology of music.

Because timbre arises from the interplay of a complex variety of sonic elements, a precise definition has eluded both music theorists and researchers in the cognitive sciences; those definitions that have been proposed will be considered here. The difficulty of devising a workable definition has, in turn, impeded both the development of a musical theory of timbre, and a generally accepted method for its description and specification.

Slawson (1985) has stated that any tractable theory of musical timbre should have a psychoacoustical component; this can equally be said of the design of tools for timbre specification. For this reason, most of this chapter is concerned with timbre as a field of psychoacoustical study and, in particular, with studies which have informed proposals for its control and manipulation (examined in the following chapter). A number of approaches are identified, and key research in each of these areas examined. Most importantly, the notion of *timbre space* is introduced and discussed, as this forms the basis of the empirical work presented in chapter five.

3.2. Terminology and definition

The term 'timbre' has become the generally accepted one in the considerable literature devoted to the subject, and will be the one used here; however, there have been a number of other suggestions, many derived from a perceived analogy with light and colour (discussed later). *Sound colour*, coined by Slawson (1985), and 'tone colour' (much used in treatises on orchestration) are direct translations of *Klangfarbe* and *Tonfarbe* respectively. Erickson's (1975) proposal of the term *clangtint* does not seem to have been taken up anywhere else.

Seashore (1967) made a distinction between the time-variant and time-invariant components of sound, drawing on an analogy with film – specifically, the illusion of movement caused by the successive display of a series of still pictures of moving objects. Thus, *sonance* was the aural sensation caused by the time-variant aspects of sound (onset, vibrato, decay, spectral fluctuation etc - the 'motion' of a sound, as it were), while the 'timbre' was equivalent to a single picture from the series, and is determined purely by its instantaneous spectrum.

A common definition of timbre is the 'quality' or 'character' of a musical instrument (Pratt and Doak, 1976). Butler's definition of timbre (Butler, 1992) views it as that which conveys the identity of the originating instrument. This raises the question of how much timbre perception is tied up with issues of identification, and is one which will be revisited later in this chapter. However, most recent studies of timbre take as their starting point the ANSI standards definition in which timbre is stated as being "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" – that is to say, timbre is what is left, once the acoustical attributes relating to pitch and loudness are accounted for. This definition, of course, raises the question of how timbral differences are to be defined in isolation from loudness and pitch when these qualities are not dissimilar. Pratt and Doak (1976) proposed refining the definition so that it read: "That attribute of auditory sensation whereby a listener can judge that two sounds are dissimilar using any criteria other than pitch, loudness or duration." Over and above the very subtractive view of timbre implied by these two definitions (when pitch and loudness are accounted for, timbre is what is left), it is also assumed that timbre is a sonic quality which is orthogonal to, and independent of the vectors of pitch and loudness. As has been noted (Krumhansl, 1989), this is not an assumption that can safely be made; it is by no means clear that judgments of timbral differences can be entirely decoupled from those of pitch, loudness or duration.

This 'three axis' model of musical sound – pitch, loudness and timbre - is nevertheless implicit in Western musical theory and notational practice: there is a separate set of conventions/symbols for the representation of pitch (the height of the symbol on the stave), for loudness (*pp*, *mp*, *mf*, *ff* etc), and for timbre (the name of the instrument, as well as symbols used to signify details of momentary

articulation). It is also reflected in the design of subtractive synthesizers, where the user is provided with 'handles' to these three nominal attributes in the form of voltage-controlled oscillators, filters and amplifiers. As early as 1911, however, Arnold Schönberg conceived timbre more holistically; in his *Theory of Harmony* (Schönberg, 1911), in which he speculates on the possibility of building musical structures based on timbre, he writes:

> 'The distinction between tone color and pitch, as it is usually expressed, I cannot accept without reservations. I think the tone becomes perceptible by virtue of tone color, of which one dimension is pitch. Tone color is, thus, the main topic, pitch a subdivision. Pitch is nothing else but tone color measured in one direction.' (p. 421)

3.3. Timbre in music

3.3.1. Historical perspective

A growing interest in the colouristic possibilities of instruments, both as solos and in groups, can be traced through the seventeenth and eighteenth centuries. However, Berlioz's *Grand traite d'instrumentation et d'orchestration modernes o*f 1844 (Macdonald, 2002) was perhaps the first to treat timbre as a distinct and independent musical element: in it, he describes the art of writing for the orchestra as

> 'the use of [the] various sonorities and their application either to colour the melody, harmony or rhythm, or to create effects *sui generis*, with or without an expressive purpose and independent of any help from the other three great musical resources… ' (p. 6)

Rimsky-Korsakov's *Principles of Orchestration* (1891), published nearly fifty years later is an illustration of the growing importance ascribed to timbre as a musical resource; in it, he says

> 'It is a great mistake to say: this composer scores well, or, that composition is well orchestrated, for orchestration is part of the very soul of the work. …. One might as well say that a picture is well drawn in colours.' (p. 2)

Walter Piston's treatise on orchestration (Piston, 1955) describes techniques by which instrumental sounds can be blended to create a single distinctive timbre. (Piston's extensive vocabulary of adjectives for sound and combinations of sounds was investigated in a 1993 study by Kendall and Carterette, which is reviewed later in this chapter.) Robert Erickson (1975) has said that 'composing should include composing the orchestra'.

By the beginning of the twentieth century, orchestral colour and texture was being used increasingly as a means of structuring and articulating musical forms and gestures. In the score of *Farben*, his study in orchestral colour from *Five pieces for Orchestra* Op.16 (1909), Arnold Schönberg (1874-1951) instructs the conductor to ensure that no single instrument makes itself conspicuous: the listener is to hear only changes in overall timbre, as different instruments drop in and out of the texture. The work of Edgar Varèse (1883-1965) shows a similar foregrounding of timbre; the register and dynamic of individual instrumental sounds in *Hyperprism* and *Octandre* (1923), for example, are carefully chosen in order to promote a perceptual fusion, such that they are perceived as a single timbre. The dense mesh of instrumental/vocal sound heard in the 'micropolyphonic' works of György Ligeti (1923-2006) such as *Lux Aeterna*, *Requiem* and *Atmospheres*, foregrounds mass, colour and texture at the expense of melodic and rhythmic elements.

3.3.2. Metaphor and analogy

The difficulty of defining timbre, or attributing it to a clearly defined set of acoustic parameters has meant that its description in music has necessarily been indirect, making use of metaphor and analogies with the senses of vision and touch. The association of sound with colour, or *Tonfarbe* (tone-colour) in German was noted by Rimsky-Korsakov (1891), but is in fact of long standing. References can be found in the writings of Ptolemy, who, in his *Harmonics,* maintains that beauty is only perceived through the two senses of sight and hearing, which cooperate with each other 'as if they were sisters' (Barker, 2000). In everyday speech, we talk of sounds as being 'bright' or 'dark', and conversely, describe vivid, garish colours as being 'loud' and duller colours as 'muted'. Other terms typically used to describe timbre are borrowed from a vocabulary of texture - 'rough', 'smooth', 'sharp', 'blunt', 'fine', 'coarse' etc. Psychoacoustical studies which have investigated the semantic link between these and similar descriptors and quantifiable acoustical attributes will be reviewed later in this chapter.

3.3.3. Theories of musical timbre

The foregrounding of timbre as a means of structuring music has not been accompanied by a development of a generally accepted musical theory of timbre (as distinct from an empirical one based on psychoacoustic experiment). This is, in part, due to problems relating to its notation and visual representation. Musical notation has been the basis on which Western music theory has modelled musical structures and gestures. Pitch, time values and dynamics are represented by, and notated in ordinal scales (C5 is higher than C4, a minim is longer than a crotchet, *ff* is louder than *mp*); such scales make explicit the relationships between pitches,

durations and dynamic levels and permit the construction of theories for their musical articulation (Hajda, Kendall *et al.*, 1997). The fact that such simple ordinal relationships do not exist for timbral categories (in what sense is a clarinet greater or less than a trombone?) is a problem for the development of a musical theory of timbre.

This problem has been tackled in two contrasting ways. Slawson (1985) starts out from the premise (stated at the beginning of this chapter) that, firstly any useful musical theory of timbre needs to have a psychoacoustic component, and secondly, should define that which is invariant in timbre – that is to say, how can loudness and pitch and other aspects of sound be changed while keeping timbre, or sound colour, fixed. Secondly, it should be able to define operations, analogous to those that can be carried out on pitch (inversion, transposition etc), which can be performed on timbre. A later study (Slawson, 1989) proposes a timbral scale made up of discrete vowel sounds arranged around a circle according to the degree of perceptual similarity between them. A timbral interval between two given vowels is then expressed in terms of the number of anticlockwise steps around the circle which separate them (the inversion of the interval is then the number of clockwise steps). From this also arises the possibility of timbral motifs which can be transposed by clockwise/anti-clockwise shifts around the circle. Slawson goes on to speculate on the possibility of timbral transposition and inversion.

The idea of timbral structures and transformations derived from pitch is also taken up by Lerdahl (1987) who poses the idea of timbral consonance and dissonance. Drawing on his work with Jackendoff (Lerdahl and Jackendoff, 1983), he has suggested ways in which musical gestures based on timbral transformation

and variation can be articulated through mechanisms similar to the grouping and prolongational structures of generative music theory.

The work described above assumes a correspondence between pitch, loudness and timbre such that methods of musical transformation and change based on pitch can be extended and applied to the other two parameters. In fact, the way in which pitch, loudness and timbre are perceptually structured differ considerably. In general, loudness is a one-dimensional vector which maps fairly directly to signal power/amplitude. Similarly, pitch perception also maps to a single dimension, but is also characterised by an equivalence relation – the octave (Balzano, 1986). Thus, a more appropriate model for pitch is a spiral rather than a straight line. Shepard (1982), in fact, proposes a five dimensional solution, taking in the chroma circle[6], pitch height and the circle of fifths). Finally, timbre does not clearly correlate with any one perceptual dimension, arising as it does from a complex interplay of a number of quite distinct sonic attributes. These qualitative differences suggest that the scope for a musical theory of timbre which draws too much on, for example, pitch theory, is limited (there is no such thing as a timbral octave, for example).

3.3.4. Classification and taxonomy.

While there is no generally agreed theoretical underpinning for the musical articulation of timbre, there have been a number of proposals for its more general classification and taxonomisation. These approaches fall broadly into two categories – the *acousmatic* and the *ecological* – although it should be noted that the boundary between the two is not clearly defined.

[6] *Chroma,* or pitch class is the property shared by a musical pitch and all other pitches that stand in octave relationship to it – thus C_1, C_2, C_3, C_4 etc all have the same chroma.

3.3.4.1. Acousmatic approach

Acousmatic music is music where 'the sources and causes of the sounds are invisible' (Smalley, 1997) - relayed through loudspeakers, for example (although, as Smalley notes, any CD recording could be regarded as acousmatic). More specifically, acousmatic sound is sound which is heard decoupled from its source. Derived from Pythagoras' method of teaching, (who, it is said, delivered his lectures to his disciples over a period of five years, while hidden behind a curtain), the term is primarily associated with Pierre Schaeffer, whose *Traité des objets musicaux i*s one of the more well known taxonomical studies (Schaeffer, 1966). The *Traité* describes a method for classifying sounds – or more specifically sound objects (*objets sonore*) – using a number of perceptual categories; this classification is realised through a process of 'reduced listening', in which the attention of the listener is disconnected from the physical origin of the sound, and is instead focussed on the directly audible attributes of the sound itself. An example category is mass (*masse*), in which sounds are grouped according to whether they are pure tones (type M1), complex pitched sounds (type M2), complex, non-variable non pitched sounds (*mass fixe*) (type M3), sounds that vary somewhat (type M4), and sounds that vary a lot (type M5). A second category is treatment (*facture*), defined as "the way in which sound is communicated or made manifest throughout its duration"; within this category, sound may be described as continuous, discontinuous or impulsive.

Slawson (1985) has observed the contradiction between the notion of reduced listening, and Schaeffer's own description of a sound as having a "mechanical allure". Hajda, Kendall, Carterette and Harshberger (1997) also criticise Schaeffer, observing that the variables used in his taxonomy were arrived

at through introspective listening rather than empirical observation and measurement, and are consequently highly subjective. Trevor Wishart (1986) notes that:

> "in our common experience, we are more often aware of the source of a sound than not, and studies of behaviour and aural physiology would suggest that our mental apparatus is predisposed to allocate sounds to their sources" (p. 129)

Nevertheless, Wishart acknowledges the work of Schaeffer and the Group de Recherches Musicales by revisiting a number of categories proposed in the *Traité*; in particular, those which describe a sound's temporal evolution. The categories he arrives at – the 'discrete', the 'iterative' and the 'continuous' (Wishart, 1986) - are based on the assumed physicality of the sound source, and essentially describe the patterns of energy disposition characteristic of the sound producing mechanism. So, for example, a single drumstroke is a 'discrete' sound, while a drum roll is 'Iterative'. A sustained flute sound, however, is 'continuous'. In this schema, sound is classified by its dynamic morphology (Wishart, 1996), which may be *intrinsic* (caused by an instantaneous input of energy, such as plucking or striking) or *imposed* (caused by a continual input of energy such as bowing, or a steady air stream).

Denis Smalley's approach to timbre taxonomy (Smalley, 1986) has much in common with that of Wishart, in that sounds are characterised by their morphology. An important difference, however, is that there is no overt linkage with physical mechanisms or gesture; the notion of the acousmatic is much more implicit in Smalley's schema. Building on a set of central reference points placed on a note-to-noise continuum – *note* (note proper, harmonic spectrum, inharmonic spectrum), *node* and *noise*, Smalley assembles a palette of *morphological archetypes* -

characteristic and recognisable patterns of spectral change over time (e.g., *attack-impulse*, *closed attack-decay*, *graduated continuant)* – whose profiles can be represented using symbols. This set can be extended and hybridised to form more elaborate morphological patterns.

3.3.4.2. Ecological approach

The ecological approach to the understanding of sound has started out from the opposite premise; that, precisely because it conveys information about both its source - size, distance, direction and speed of motion – and its environment, sound (in general) cannot be perceptually disassociated from the materials and actions which generate it. The most important proponent of this view is William Gaver, who draws on Gibson's ecological approach to visual perception (Gibson, 1966; Gibson, 1979) in proposing an alternative explanatory framework (Gaver, 1993).

This view of sound is one in which we understand and manipulate sounds along dimensions representing the attributes of the sound sources rather than those of the sounds themselves. A given sound provides information about an *interaction of materials* at a *location* in an *environment*. The attributes of the sound source describe the *interaction*, the material (restoring force, density, damping, homogeneity) and the *configuration* (its shape, size, nature of support). All of these attributes have an effect on the amplitude, spectrum, bandwidth of the sound as well as its evolution with respect to time. This level of description provides a framework for describing sonically complex events – the crumpling of paper, breaking, bouncing, sawing wood etc. Gaver's work has informed a number of proposals for the design of auditory icons (to be reviewed in the next chapter).

3.4. Acoustical and psychoacoustical studies

3.4.1. Introduction

We turn now to consider empirical psychoacoustical research into the nature of timbre, conducted over the past forty years. A number of distinctly different approaches have emerged; these were identified and reviewed in a 1997 analysis of methodologies (Hajda, Kendall *et al.*, 1997), and are summarised here.

3.4.1.1. Identification

The verbal labelling of a stimulus by class or category instrument. In studies of this type, subjects are asked to name the instrument category which most closely resembles the stimulus (Clark, Robertson *et al.*, 1964; Saldanha and Corso, 1964; Wedin and Goude, 1972).

3.4.1.2. Categorisation and matching

The categorisation methodology requires subjects to state which of a set of stimuli (the 'choice' set) most closely resembles a given member of a 'model' set (Kendall, Carterette *et al.*, 1995).

3.4.1.3. Verbal attributes.

These studies have sought to identify correlations and associations between attributes of sound and the adjectives/adverbs used to describe them (Lichte, 1941; Bismarck, 1974; Kendall and Carterette, 1993). These studies will be reviewed in greater detail in this chapter.

3.4.1.4. Proximity rating

This approach has been successfully applied by a number of researchers (Grey, 1975; Grey, 1977; Iverson and Krumhansl, 1993; Hourdin, Charbonneau *et al.*, 1997; Kaminskyj, 1999). In studies of this type, subjects are presented with pairs of tones and asked to rate them on a numeric scale for similarity. Typically, the matrix of data generated is analysed by a multidimensional scaling (MDS) algorithm, and a coordinate space of low dimensionality is generated, containing a number of points, each representing one of the stimuli, disposed in the space such that the distance between any two reflects the degree of perceived similarity. Again, studies of this type will be reviewed later in this chapter.

3.4.1.5. Discrimination

In general, this approach has been used to establish just noticeable differences between stimuli, and has been less used in timbre investigations, although it has been employed in a study designed to establish whether original recorded tones and signals in which these tones were subject to various levels of spectrotemporal simplification could be discriminated (Grey and Moorer, 1977).

3.4.1.6. Timbre perception versus identification

What emerges from these studies of timbre, and which has a bearing on some of the timbre specification methods described in chapter four of this thesis, is that *timbral perception* is distinct from *timbral identification*; we are able to hear timbral change in the sound of a violin, for example, while at the same time still identifying the sound as that of a violin. One of the important and stated objectives of psychoacoustic research in this area is to identify salient attributes of

sound which contribute to timbral change; another is to locate those invariances in the acoustic wave which enable us correctly to identify a particular sound source in a variety of different acoustical contexts. While these are complementary objectives, they are nevertheless distinct. In the same way that a photograph of a person remains recognisable even when tinted, many sounds retain a distinct aural identity even when subjected to a degree of acoustic modification. A cello or a saxophone is recognizable as such, even in a poor recording or in a reverberant room (Risset and Wessel, 1999). At the same time, we can clearly hear that a timbral change has taken place. It could be said that while timbre perception relates to a multidimensional continuum of acoustic parameters, timbre identification maps onto a more granular and discrete space; within this space, there are tolerances within which sonic parameters can change without affecting the identification of a given sound.

Donnadieu (2007) has observed that 'the concept of timbre is much more general than the ability to distinguish instruments', noting that a timbre may be, variously, one single instrumental sound, or the gamut of sounds produced by that instrument; alternatively, it may refer to a combination of individual sounds, or to hybrids or chimerae for which there is no known sound source. For the purposes of this thesis, these distinctions are important, as a synthesizer user may wish to create a sound which falls into any of these categories, and a useful UI should be capable of providing the means of doing this.

With this in mind, this section reviews the psychoacoustic research literature on timbre, beginning with its frequency domain, 'spectral' aspect, before going on to consider its dynamic time-variant component, at both the macro level (overall dynamic envelope) and the micro level (the spectrotemporal fluctuations which a sound undergoes, and which have been shown to be perceptually important).

Investigations of mappings between verbal descriptors of sound and measurable acoustical attributes are also reviewed here in some detail. Finally, and most importantly for this study, the notion of a timbre space, in which a sound is conceptually located in an *n*-dimensional space of sonic attributes, is examined. Psychoacoustical work which takes this particular abstraction as a starting point is reviewed in detail, and a distinction is made between *perceptual* and *attribute* spaces.

3.4.2. Frequency spectrum

The association of timbre with the frequency spectrum of the steady state portion of an instrumental tone was first made by Ohm in 1843, and elaborated on by Helmholtz (Helmholtz, 1954). Ohm's acoustical law states that the timbre of a musical sound is attributable to the pattern of amplitudes of those harmonics. Both Ohm and Helmholtz, however, maintained that 'differences in musical quality of tone depends solely on the presence and strength of partial tones, and in no respect on the difference in phase under which these partial tones enter into composition' – that the ear is, in effect, phase deaf. Later studies have questioned this, however, and have shown that a shift in the phase relationships of the component harmonics can be perceived, albeit weakly (Mathes and Miller, 1947; Plomp and Steeneken, 1969).

Nevertheless, a connection between timbre and frequency spectrum can generally be made. The 'brightness' of a sound can be modified by boosting or attenuating the amplitudes of its higher partials; the 'presence' of a sound seems to be associated with the amplitude of the spectrum around 2000 Hz (Risset and Wessel, 1999). This simple model, however, is less useful as a means of explaining of *invariances* in a given sound. The shifting of a spectrum up and down in

frequency, preserving the amplitude and frequency ratios between its partials, nevertheless results in changes in timbre. (This can be demonstrated if a tape recording of, say, a clarinet, or normal speech is played back at double speed. In fact, the timbre of even a sine wave will noticeably change with frequency (Köhler, 1915; Stumpf, 1926).

Consideration of a sound's *formant* characteristics provides a better model for understanding the relationship between spectrum and timbre. A formant is a peak in the spectral envelope of a sound which is often associated (particularly in the case of vocal sounds) with resonances in the sound source (Risset and Wessel, 1999). Changes in the frequency of the fundamental do not result in a shift in the formant frequencies – thus sounds of differing pitch originating from a given sound source will have correspondingly different spectral envelopes. Slawson (1968) and subsequently Plomp & Steeneken (1971) demonstrated that perceived timbral similarities were more easily attributable to invariances in the formant structure than to invariances in spectrum envelope. (There does seem to be a limit to the pitch range in which this association can be said to operate – a study (Handel and Erickson, 2001) found that listeners were unable to say whether two wind instrument notes, separated by an octave or more, were played by the same instrument or on two different ones; they were similarly unable to say whether a vowel sung at different pitches, again separated by an octave or more, was sung by the identical or a different soprano or mezzo-soprano).

3.4.3. Temporal characteristics of sound

That timbre perception is not solely attributable to the characteristics of the steady state stage of a tone has been known since Helmholtz' time; most sounds are not time invariant, and the way they evolve in time, both spectrally and in

their overall dynamic envelope, provides important cues for identification. Because the empirical work presented later in this thesis makes use of sounds which are, for the most part time-invariant (i.e. steady state), the review of the research literature concerned with the temporal aspects of timbre will be more condensed than that of other aspects of timbre perception; its importance, however, is acknowledged and emphasised.

The onset and offset characteristics of a sound – how it begins and ends - are important identification cues; in fact, the removal of the initial attack segment of a note significantly impairs subjects' ability to recognise the source (Stumpf, 1926). Similarly, subjects also have difficulty in identifying a note if its overall amplitude envelope is reversed – that is to say, if it is played backwards (George, 1954). Richardson (1954), in a study which looked at the amplitude envelope of individual harmonics in the onset phase of organ pipe tones, speculated that the importance of this phase for identification was comparable to that of the steady state spectrum.

Berger's study of instrumental timbre (1964) examined the relative perceptual salience of spectrum and the onsets and offset characteristics of notes played on a number of wind and brass instruments. Recorded tones were presented to listeners i) unedited, ii) with the onsets and offsets removed, iii) unedited but reversed and iv) unedited but heavily low pass filtered; listeners were asked to identify the instrument in each case. Not surprisingly, the percentage of correct identifications was highest for the unedited stimuli (59%), and was significantly lower in the case of reversed stimuli (42%); removal of onsets and offsets reduced this figure further to 35%. The filtered stimuli were

recognised by only 18% on average[7]. Overall, these results suggest that the temporal and transient characteristics of a sound, while important for the purposes of recognition (at least of musical instruments) are less salient than its spectral properties.

A more detailed study of onset and offset saliences in musical instrument identification was carried out by Saldanha and Corso (1964). A wider range of instrumental types were used here (strings in addition to woodwind and brass); each instrument was played with and without vibrato, and at three different pitches (F4, C4 and A4). For each of these cases, five types of edit were created - i) a tone with initial transients and shortened steady state, ii) entire tone with shortened steady state, iii) entire unedited tone, iv) shortened steady state only, and v) shortened steady state and final transients. The results suggested that those stimuli where the initial transient had been preserved were most easily recognised (47%), and those which consisted only of steady state, or steady state and offset, least well recognised (32% in both cases). (Curiously, unedited tones were less frequently correctly identified than tones which consisted of a shortened steady state.) As in Berger's study, it was noted that there was a degree of confusion between instruments of the same family, and that some instruments were more easily identified than others – although, in Berger's study, the flute was one of the instruments least often recognised, whereas in Saldanha and Corso, it emerged as one of the most recognisable. Other conclusions of the study were that pitch affected identification (more correct identifications were made at F4 than C4 or A4), and that a vibrato tone is better identified than a non-vibrato tone. Saldanha

[7] These percentages are surprisingly low, particularly in the case of unedited stimuli. However, as Berger notes, a number of instruments belonged to the same instrumental family (cornet and trumpet, for example, or alto and tenor saxophones), and listeners could therefore be easily confused. Furthermore, because of the need to equate frequency for the purposes of the test, several instruments were playing out of their normal ranges, and were therefore less easily recognised. The results also showed variation according to instrument - the oboe was most easily recognised under all conditions except the filtered case, and the flute and trumpet were most difficult to recognise.

and Corso surmised that the order in which partials appear and disappear in the onset and offset stages may provide an important identification cue.

Further discussion of the spectrotemporal aspects will be presented as part of the review of multidimensional scaling studies and timbre space later in this chapter.

3.4.4. Timbre and language

The section on 'musical' timbre (section 3.3) looked at the lexicon of descriptors commonly used to describe musical sound; since the 1940s, there have been a number of studies designed to identify correspondences between the acoustical attributes of (for the most part) steady state tones and the adjectives used to describe those tones. Some of these studies (Lichte, 1941; Bismarck, 1974) have focussed on one particular term (e.g., *roughness* or *sharpness*) and looked at the different attributes which contribute to this single perception. Lichte used electronically generated complex steady-state tones as stimuli, the amplitudes of whose harmonics variously increased or decreased linearly with frequency or exhibited a peak or trough at the eighth and ninth harmonics. Pairs of tone were played to subjects, who were asked to say whether the second stimulus of the pair was *brighter* or *duller* than the first. The study concluded that the perception of *brightness* was related to the location in the spectrum of the mid-point of the energy distribution (the spectral centroid). Another set of stimuli varied the amplitudes of the odd numbered harmonics relative to those of the even harmonics. Again, these were presented to subjects in pairs and subjects were asked to state whether the second of each pair was *thinner* or *fuller* than the first. *Fullness* seemed to be associated with spectra where the odd harmonics were of greater amplitude than the even ones.

3.4.4.1. Semantic differential

A methodology frequently employed in subsequent studies is the *semantic differential* method (Osgood, Suci *et al.*, 1957); used in sociological research, it is a measurement tool in which subjects are asked to rate a particular concept or stimulus on a series of seven point bi-polar semantic scales e.g., *heavy-light*, *wise-foolish* etc. Such scales together form a multidimensional 'semantic space', which can be analysed by means of (for example) Principal Component Analysis (PCA) in order to determine underlying variables which contribute to this perception. Thus, in psychoacoustic studies, typical scales used are *calm/restless*, *light/dark*, *rich/poor*, *solid/hollow* etc, which can be shown to map variously to spectral centroid, bandwidth etc. Common to many of these studies is an initial phase in which a number of subjects freely volunteer timbral adjectives; those which occur most often in subject responses are then subsequently used in the semantic scales for rating sound stimuli.

The semantic differential methodology was first used in a study of the timbral vocabulary used by US Navy sonarmen to describe sonar signals (Solomon, 1958; Solomon, 1959). The adjectives chosen were semantically quite diverse, drawing on colour analogies (*green/red*, *colorful/colorless*, *dark/bright*), on emotional associations (*happy/sad*, *calming/exciting*), aesthetic (*beautiful/ugly*) as well as employing some quite high level and abstract descriptors (*good/bad*, *definite/uncertain*, *obvious/subtle*, *masculine/feminine*). Factor analysis revealed that these scales clustered into seven factors for which interpretation could be made, accounting for 40% of the total variance. Subsequent analysis showed strong positive correlation between the strength of energy in the lower frequencies and ratings of *heavy*, *large*, *wide* etc; and a strong correlation between energy in the

higher frequencies and ratings of *light*, *small*, *narrow* etc. Sounds with energy concentration in the 600-1200 Hz region were judged to be *tight*. Interestingly, the '*strange/familiar'* scale was associated with energy at 75-150 Hz and 300-600 Hz at the *strange* end, and with energy at 4800-9600 Hz at the *familiar* end; similarly, low frequency energy (75-150 Hz) was associated with *colorful* ratings whereas high frequencies were associated with *colorless* ratings.

In von Bismarck's frequently cited study (Bismarck, 1974), the range of stimuli used was larger and more varied than that used in Solomon's work. Thirty scales were used (e.g., *gentle/violent*, *rounded/angular*, *dull/sharp* etc - like Solomon, von Bismarck selected adjectives drawn from a variety of domain vocabularies – colour, texture, aesthetic, emotional. The frequency spectra of the thirty-five time-invariant sound samples used in this study were selected in order to represent the most prominent characteristics of instrumental sounds and those of voiced and unvoiced speech sounds. In addition, sounds were included for their overall spectral envelope (-6 dB per octave, -12 dB per octave etc), and for the timbral effect of prominent odd and even harmonics.[8] In the first phase of the study, sixteen subjects (divided into musicians and non-musicians) were firstly asked to rate thirty five steady state sounds on a set of two bi-polar scales (*dark/bright* and *rough/smooth*). In general, there was found to be good agreement between the ratings of musicians and non-musicians. Secondly, subjects rated a single repeated sound on a set of thirty scales (e.g., *dull/sharp*, *relaxed/tense*, *solid/hollow*). Factor analysis showed that 91% of the variance could be accounted for by just four factors: the first one represented attributes such as *sharp*, *hard* and *loud*; the second one was characterised by *compact*, *boring* and *narrow* ; the third by *full* and the fourth by *colorless*.

[8] A timbral space broadly characterised by these attributes is used in the empirical work presented in this thesis, described in chapter six.

It was noted, however, that the value of these factors in describing timbre was limited by significant variation between subjects' individual ratings; only scales such as *round/angular* and *reserved/obtrusive* were used with a degree of unanimity, and only the first factor – represented by *sharpness* - seemed to be a psychoacoustically useful semantic scale.

A closer examination of the perception of *sharpness* and its corresponding acoustical attributes for a number of different types of steady state timbres was also done – this revealed that *sharpness* appeared to be related to the upper and lower limiting frequencies and to the slope of the spectral envelope. As von Bismarck noted, these results were consistent with the findings of Plomp (1970) (timbre as multidimensional attribute of complex tones) and Plomp & Steeneken (1971) (pitch versus timbre); namely, that the absolute position of the spectral envelope was more salient to timbre perception that its position relative to the fundamental (see discussion above).

A study conducted at around the same time as that of von Bismarck identified three semantic scales - *dull/brilliant*, *cold-warm*, *pure-rich* as being meaningful, based on a questionnaire given to music students (Pratt and Doak, 1976). Listening tests run, using electronically generated steady state spectra as stimuli, showed that subjects were able to differentiate between sounds most effectively on the *dull/brilliant* scale (results compatible with those of von Bismarck), but that these three scales were not perceptually completely independent.

3.4.4.2. Verbal attribute magnitude estimates (VAME)

The usefulness of the semantic differential methodology in the study of audio perception has been questioned (Kendall and Carterette, 1993). While many of the studies described above have drawn on freely volunteered terms for sound description, the choice of terms (*heavy-light, dull-sharp*) for the opposite ends of a given semantic scale is not necessarily that which would be spontaneously chosen by subjects (Donnadieu, 2007). Nor are the terms at each end necessarily antonymous (is *dull* the opposite of *sharp*?). An alternative approach is one which make use of verbal attribute magnitude estimates (VAME), in which subjects are asked to rate on a sliding scale the degree to which an adjective describes the stimulus, where one end of the scale is (for example) *sharp*, and the other is its negation, *not sharp*. This method was adopted in an investigation of von Bismarck's adjectives (Kendall and Carterette, 1993), which was significantly more successful in differentiating ratings by instrument, and generating a principal component analysis.

3.4.4.3. Other languages

Issues of cultural specificity are inevitably raised by studies of this type where the vocabulary used is in a language other than English (Faure, McAdams *et al.*, 1996; Moravec and Stepánek, 2003). The first of these studies used twenty-three VAME scales derived from a French vocabulary – e.g., *pincé* (plucked), *sec* (dry), *large* (wide) - such as 'not very metallic' and 'very metallic', using twelve synthesized sounds, some of which imitated standard Western instruments (e.g., trumpet, clarinet etc). They found significant agreement in judgements between subjects, and correlations between the positions of sounds on each dimension of a four dimensional multidimensional scaling space. Terms like *pincé* and *soufflé*

(blown) were clearly related to the 'attack' dimension, whereas *aigu* (sharp or shrill), *haut* (high), *grave* (low/deep) and *bas* (low) could be located on the spectral centroid dimension. Only one verbal attribute – *riche* (rich) - could be traced to the third dimension, spectral fine structure, the ratio of energy between even and odd harmonics, and none with the fourth. Subjects were also asked which of these terms was most relevant for describing the sounds – *attaque* (attack), *doux* (soft), *sourd* (muffled or dull) and *métallique* emerge as the most applicable .

A Swedish study of saxophone timbre (Nykänen and Johannsen, 2003) took a slightly different approach. The aim of this investigation was to find correlations between Swedish adjectives frequently used by saxophone players, the use of vowel-similes, and acoustically measurable characteristics of the saxophone tone. Following the practice of a number of other studies, a number of terms were identified through initial interviews with Swedish saxophonists. The subjects were in two groups – one of sax players, the others of experienced listeners who did not play. Subjects were asked to rate stimuli according to these descriptors (using a VAME methodology) and were also asked to estimate the extent to which the stimuli were best described by a number of vowel sounds. The stimuli were played by two different saxophonists, each on two different saxophones – so there were four versions of each stimulus. Each sound stimulus was measured and classified according to fundamental frequency, whether the spectrum had prominent formants, formant frequencies, overall spectral bandwidth, loudness, roughness (Aures, 1985), sharpness (Bismarck, 1974) etc. Principal Component Analysis revealed *sharpness* and *roughness* to be two important factors, but also the frequencies of a number of the formants. Of interest to this thesis are the observations on the possible importance of formant frequencies, as one of the timbre spaces used in the empirical work is

characterised by three formant centre frequencies; Nykanen *et al* note that there is a clear difference between the timbres of sounds produced by the two players.

Kendall and Carterette's investigation of the adjectives used by von Bismarck (which were in German), suggested that any correlations identified in one language may not necessarily be valid in another (Kendall and Carterette, 1993). Simultaneous dyads of woodwind instrument notes were used as stimuli, and subjects asked to rate them on the same eight semantic scales identified by von Bismarck. They concluded that these scales failed to differentiate between the stimuli used. As the authors noted, this may be simply be attributable to the stimuli themselves (which were different from those used by von Bismarck), but could equally be because of mismatched translation (does *sharp* in English mean the same as *scharf* in German when applied to sound?).

3.4.4.4. Discussion

The use of verbal directives for, and mappings to synthesis parameters will be discussed in the following chapter; however, the psychoacoustical research literature reviewed here throws up a number of issues for the design of such systems.

Firstly, the use of VAME methodology has been shown to be of greater use than the standard semantic differential method in understanding the dimensionality of verbal spaces and their correspondances with acoustical attributes.

Secondly, there is the issue of the degree of common understanding of a particular descriptor. While some studies have found broad agreement as to what

a particular term 'means' (Faure, McAdams *et al.*, 1996), other studies have reported significant variation. A recent study (Darke, 2005) divided twenty-two subjects into two groups; using a VAME methodology, each subject was asked to rate fifteen instrumental sounds for *brightness*, *harshness*, *brassiness* etc. Significant differences were found between the two groups; the author notes that it is 'not clear whether the subjects are rating the instrument rather than the sound', and while methodological issues may have affected the results in some cases, there was no strong evidence that subjects agree on how to verbally communicate timbral descriptions. This was also found to be the case in a study of timbral adjectives applied to the pipe organ (Disley and Howard, 2003) ; while there seemed to be common understanding for terms like *bright* and *warm*, other terms like *full* and *balanced* elicited a wide variety of different responses (although, interestingly, agreement on these adjectives increased with the degree of musical training). Descriptors like, for example, *loud* are ambiguous; both Bismarck (1974b) and Kendall and Carterette (1993) note that, although the stimuli used had been equalised for loudness, *loud* was nevertheless an important differentiating factor; loud sound seems not to be perceptually the same thing as loud timbre.

Thirdly, the assumption made in a number of studies (Bismarck, 1974; Pratt and Doak, 1976) is that adjectives and semantic scales which are close in meaning can be regarded as synonymous, and can be eliminated in order to reduce the number of semantic scales[9]. This methodological practice has been critiqued (Kendall and Carterette, 1991); it is by no means self-evident, that, for example, *soothing-exciting* is semantically identical with *calm-restless*, or would be regarded as such by most subjects. The reverse is also the case, of course, as shown in Pratt *et al* – two or more semantic scales may not necessarily be perceptually independent. The naming of dimensions in PCA or MDS configurations is also

[9] Pratt and Doak, however , do acknowledge that such elimination could be seen as arbitrary and idiosyncratic.

potentially arbitrary, based on not much more than intuition (Kendall and Carterette, 1993) (this particular study identified a 'plangent' factor in a principal component analysis of adjectives used by Walter Piston, which is not a term which appears anywhere else in the literature.)

Fourthly, the choice of adjectives may vary according to the type of listener. An investigation into the terms used for describing timbre in Czech (Moravec and Stepánek, 2003) involved an initial vocabulary gathering exercise in which the subjects were musicians. The frequency of terms used was analysed by the type of instrument played by the subject, and significant differences were found; thus, for example, a term like 'shining' (English translation) was less likely to be used by keyboard players than by wind players.

Kendall and Carterette's study of the adjectives used by von Bismarck put a question mark over the cross-cultural validity of the adjectives used in the studies discussed here.

Finally, the mapping of the sound space formed by a sound's acoustical attributes to the verbal space formed by semantic scaling is, as has been noted (Kendall and Carterette, 1991), almost certainly not linear, and many different mappings and sub-set spaces may be possible for sounds whose envelopes are impulsive (e.g., xylophone) or non-impulsive (e.g., bowed violin). For example, a sound or set of sounds which is vowel-like, having a distinct formant structure, may invoke a particular class of descriptors which are irrelevant or less useful in describing the sound of a guitar.

3.4.5. Multidimensionality and timbre space

3.4.5.1. Introduction

One approach to timbre study has been to construct timbre spaces: coordinate spaces whose axes correspond to well-ordered, perceptually salient sonic attributes. This particular model, in which sounds occupy an *n*-dimensional co-ordinate space, can be traced back to Licklider (1951), and Plomp (1970).

Before we consider existing research which uses this model, there is an important distinction to be made. Individual sounds in a timbre space can be presented as points whose distances from each other either reflect and arise from similarity/dissimilarity judgments made in listening tests (Risset and Wessel, 1999), or, alternatively, where the space is the *a priori* arbitrary choice of the analyst, where the distances between points reflect calculated (as distinct from perceptual) differences derived from, for example, spectral analysis (Plomp, 1976). There seems to be no generally accepted naming convention for these two types of space; Nicol has proposed *human timbre spaces* for those spaces derived from listening tests, and *automated timbre spaces* for those which are generated solely from acoustic parameters (Nicol, 2005). However, it is proposed here to use the term *perceptual space* instead of 'human timbre space' and *feature space* instead of 'automated timbre space' as being more meaningful. In either case, the axes will be vectors representing measurable attributes of the sounds inhabiting the space.

An early study (Plomp, 1976) indicated that feature spaces can map to perceptual spaces. Single cycles of waveforms from a number of musical instrument recordings – violin, viola, cello, oboe etc – were electronically extended

and presented to subjects in groups of three. Subjects were asked to select in each triad the pair that were most similar and the pair that were most dissimilar. The data was then drawn up as a half-matrix of dissimilarity indices. The stimuli spectra were analysed using a set of 15 1/3-octave band filters, and the differences between the sound spectra of the stimuli presented in two ways. In the first, each spectrum was plotted in an *m*-dimensional space (an attribute space), where *m* was the number of 1/3 octave frequency bands. The position along each axis was then the SPL level of the corresponding frequency band, and the distance between two points a measure of the difference in spectra (the Euclidean solution). In the second, the area between the two spectrum envelopes determined the difference. In both cases, Plomp found good correlation between $D_{i,j}$ (the difference in frequency spectrum between the tones *i* and *j*) and the dissimilarity indices (as can be seen in figure 3.1, which has been derived from Plomp's data) showing that the Euclidean distances between sounds in this space are broadly proportional to the perceptual distances.

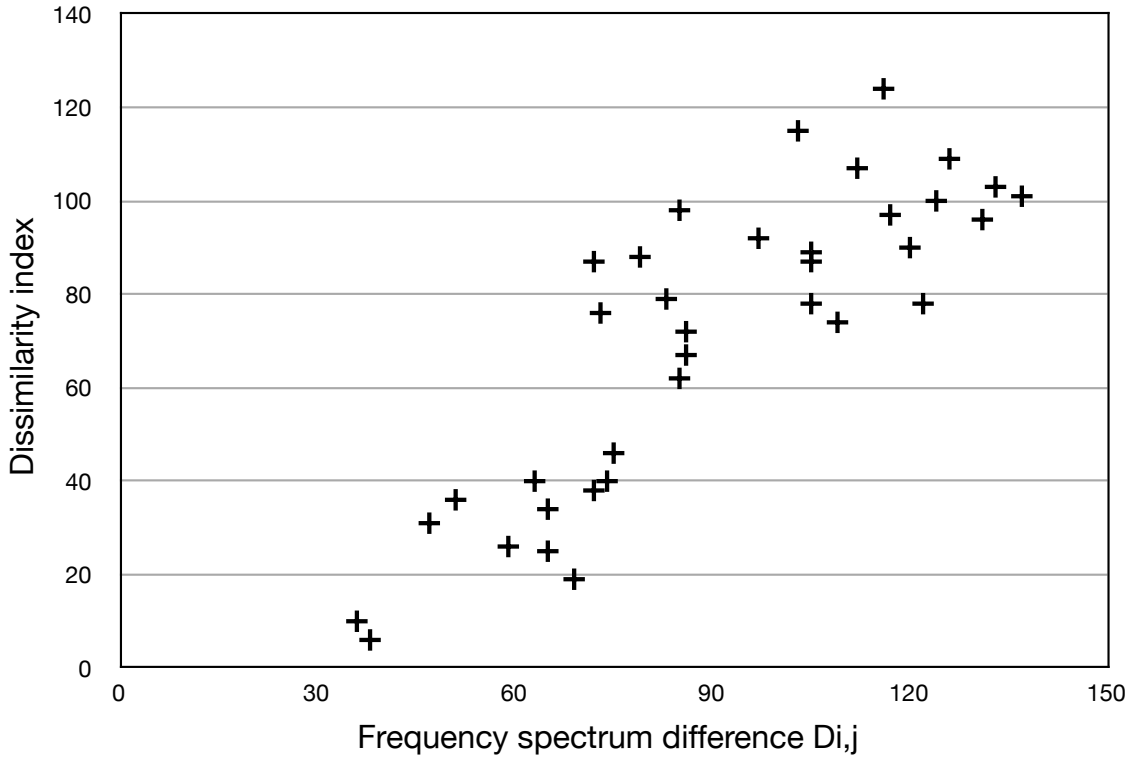


Figure 3.1: Plot of dissimilarity indices against spectral differences for nine tones adapted from musical instruments (adapted from Plomp 1970, 1976).

3.4.5.2. Multidimensional scaling studies

More recent studies have made use of multidimensional scaling to derive the axes of a timbre space empirically from data gained from listening tests (i.e. an attribute space). Because of the importance of this technique in the empirical work presented in this thesis, the principles of the technique are described here.

Multidimensional scaling, or MDS, is a set of techniques, used in (for example) psychology, sociology and anthropology, for uncovering and exploring the hidden structure of relationships between a number of objects of interest (Kruskal, 1964; Kruskal and Wish, 1978). The input to MDS is typically a matrix of 'proximities' between such a set of objects. These may be actual proximities (such as the distances between cities) or may represent people's similarity-dissimilarity judgments acquired through a structured survey or exposure to a set of paired stimuli. The output is a geometric configuration of points, each representing a single object in the set, such that their disposition in the space, typically in a two or three dimensional space, approximates their proximity relationships. The axes of such a space can then be inspected to ascertain the nature of the variables underlying these judgments. In timbre research, it has been used to identify those acoustic attributes which are salient to similarity-dissimilarity judgments of sound stimuli.

An early study (Wedin and Goude, 1972) accounted for perceptual similarities within a set of common orchestral instruments using a three dimensional model; the first factor being the general level of spectral components, the second the successively decreasing intensity of the upper partials, and the third being low fundamental intensity and increasing intensity of the lower partials.

Miller and Carterette's 1975 study identified a three dimensional solution from stimuli whose spectra contained seven harmonics above a fundamental of 200, 400 or 800 Hz, and whose amplitude envelopes were derived from the horn, the plucked string or was trapezoid. Another three dimensional solution was found where the stimuli all had the same first harmonic, but varied in the number of harmonics, individual harmonic envelopes and onset rates (Miller and Carterette, 1975). In both cases, the solution supported the hypothesis that subjects were able to scale stimuli along the dimensions which actually defined those stimuli.

Grey's subsequent and seminal study in evaluating perceptual relationships between musical instrument tones (Grey, 1977) again used MDS techniques on data derived from similarity-dissimilarity listening tests, and identified a three dimensional solution, one of whose axes could be attributed to spectral energy distribution (narrow spectral bandwidth and concentration of energy at low frequency end vs. wide spectral bandwidth and less of a concentration of energy at low frequencies). The other two axes related to time varying qualities (synchronicity in the collective attacks and decays of the upper harmonics, and the amount of high frequency, low amplitude energy in the attack segment). These findings were reproduced by Krumhansl, whose use of an extended scaling model (Winsberg and Carroll, 1988) allowed additional attributes specific to individual timbres to be revealed which improved the fit of the data (Krumhansl, 1989). MDS techniques have also been applied to the perception of simultaneous timbres (Sandell, 1989; Sandell, 1989; Kendall and Carterette, 1991) and to the dynamic time-variant attributes of timbre (Iverson and Krumhansl, 1993). Another study compared mappings of a set of synthesized stimuli generated by a Kohonen self-organising map algorithm and a perceptual matrix

derived from similarity ratings acquired from listening tests, and found significant correlation (Toiviainen, Kaipainen *et al.*, 1995).

3.4.5.3. Discussion

While MDS provides valuable data for informing theories for the 'salient dimensions or features of classes of sounds' (Grey, 1977), such data, of course, is in itself insufficient as a basis for a search strategy which would aid the selection of a desired timbre from a previously generated perceptual space. This is because the scaling solutions identify perceptually salient features of the sound, but do not necessarily define a given sound such that it could be re-synthesised from this data alone. This has been noted in a number of studies (Krumhansl, 1989), (McAdams, 1999), (Gounaropoulos and Johnson, 2006); a sound in the MDS space may have perceptually important features that no other sounds in the same space have – and, by the same token, two sounds could occupy the same location in a given MDS perceptual space, and nevertheless be audibly different. While the possibility of such features, or 'specificities', are allowed for in the EXSCAL analysis described by Krumhansl, a simple three or four dimensional MDS solution does not, of itself, work as a search space for synthesis purposes. A more recent study (Caclin, McAdams *et al.*, 2005) sought to exclude specificities by constructing a feature space inhabited by synthetic sounds which varied only by those dimensions identified by previous MDS experiments; the study confirmed the validity of the MDS timbre space model in that, overall, there was a match between the physical space and the resulting perceptual space. (This particular feature space is one of those used in the empirical work presented in this thesis, and will be discussed in chapter five.).

Clearly, however, a set of sounds which can be completely described by a three or four dimensional coordinate space is unlikely to be sufficiently complex to be musically useful or interesting. That a simple perceptual space is firstly, stable and secondly, can have predictive as well as descriptive power, however, has been demonstrated, and this makes such spaces interesting for the purposes of simple synthesis. In a study conducted by Wessel and Krumhansl, a set of timbres derived from traditional instruments – oboe, trombone, clarinet and vibraphone – was combined with a number of hybrid timbres derived from combinations of those instruments - a 'vibrone' (vibraphone and trombone), a 'guitarnet' (guitar and clarinet) and a 'trumpar' (trumpet and guitar) (Krumhansl, 1989) . A three-dimensional MDS solution was obtained in which the hybrid timbres fell into locations which were between those of the instruments which they comprised – that is to say, the vibrone was located between the vibraphone and trombone, and the guitarnet between the trumpet and guitar.

Exchanging acoustical features of sounds located in an MDS spatial solution can cause those sounds to trade places in a new MDS solution (Grey and Gordon, 1978). This study made use of the sixteen synthetic tones used in Grey's original study (previously reviewed). Eight of these tones were modified in four pairs; for each pair (trombone/trumpet, for example) the spectral envelopes of the tones were exchanged. This new set of stimuli was used as the basis for similarity/dissimilarity listening tests as before, and a new MDS solution obtained. The interpretation of the three axes of this space was the same as that for the previous study – i.e. spectral energy distribution, synchronicity in the collective attacks and decays of the upper harmonics and the amount of high frequency, low amplitude energy in the attack segment. However, what was significant in this new space is that the tone pairs had reversed positions on the spectral energy distribution axis.

Of particular interest is the suggestion that timbre can be 'transposed' in a manner which, historically, has been a common compositional technique applied to pitch (Ehresman and Wessel, 1978). In this study, two sounds, A and B, occupied points in timbre space, with AB representing the trajectory between them. Similarly, sounds C and D occupied points such that CD was parallel in timbre space to AB. Listeners were asked to rank four sounds, D1, D2, D3 and D4, in the order that they felt the trajectories C'D1, C'D2, C'D3 and C'D4 were appropriate analogies for AB (CD being the ideal solution). Analysis of the results suggested an overall preference for a ranking in which reflected the degree of parallelism to AB. These findings were reinforced by a subsequent study (McAdams and Cunible, 1992), which found evidence for the proposition that abstract timbral relationships of this type between complex sounds could be perceived both by musicians and non-musicians.

3.5. Conclusion

The mapping of frequency to pitch and amplitude to loudness is largely straightforward and well understood, presenting no problems for its representation and specification. That which is 'left over', however, when those two elements are accounted for, and which we call 'timbre' is far more elusive and intangible. Those theories that exist for the musical use of timbre have been based on analogy with theories governing the use of (primarily) pitch, proposing frameworks in which musical structure and gesture are variously articulated through timbral transposition, inversion and the ebb and flow of timbral consonance and dissonance.

What emerges from these varying approaches to the study of timbre is the question of the extent to which our perception of sounds and their degree of similarity/dissimilarity is shaped by our conscious or latent tendency to relate them to a physical source. Timbral invariance in the time-invariant spectral component of sound ('sound color'), for example, is better explained by the formant structure of the spectrum than by the spectral envelope itself – that is to say, to the invariances caused by the physical attributes of the sound source. Similarly, the effect of removing the onset transient of a musical tone is to impair substantially the ability to identify it. (The ecological approach, of course, makes the association between sound and source overt.)

Where no physical source is apparent or can be inferred, the language used to describe sound often assumes or implies one (*brassy* , *metallic*, *explosive*, n*asal*). Terms borrowed from a vocabulary of colour, texture and emotion are also typically used. The problem is that many of these descriptors are too high level to be useful; relatively few have been successfully mapped to acoustical correlates. Even where a mapping can be identified, questions of common understanding of terms across different linguistic cultures, and across different constituencies of listener, present themselves.

The other approach discussed in this chapter is that of multidimensional scaling. These have been shown to be useful for identifying salient acoustic attributes which determine similarity-dissimilarity judgments . Timbre is multi-dimensional and MDS can successfully present the relationships between a selection of sounds in a space of lower dimensionality. While such spaces clearly do not describe all the variance between a group of sounds, they can be shown to be stable and predictive; where this is the case, and where the attributes

represented by the axes of the space are well ordered and controllable, a timbre can provide a useful vehicle for synthesis.

The next chapter revisits the issues presented in this chapter, in the context of different approaches to timbre specification and reviews the current literature in the area.

Chapter 4 - Current approaches to timbre specification

4.1. Introduction.

Chapter two of this thesis discussed the usability of current synthesizer designs from an HCI perspective, and showed that effective specification and control of the timbral element of sound was hampered by the gulf between *task language* and *core language.* Chapter three explored the reasons for this, by reviewing timbre studies from a number of perspectives; musical, acousmatic, ecological and most importantly, psychoacoustic.

This chapter looks at a number of recent approaches to the design of effective means of specifying sound which bridge the user/system language gap discussed in chapter two and which have been informed by the some of the ecological and psychoacoustic studies examined in chapter three. Thus, physical modelling – sound as the output of the interaction between a number of software 'acoustic' components – can be seen as the application of the ecological approach, in that the physical origins of sound are made explicit. Other interfaces for timbre specification make use of verbal directives; and the notion of timbre space has formed the basis of a number of timbre specification proposals to be reviewed here.

Techniques that have been used in these studies vary considerably. In order to provide context for the empirical work presented in the next three chapters, they will be discussed in two broad categories. The first category will include those studies which have treated timbral multidimensionality as essentially a data reduction problem, and which have proposed spaces of reduced dimensionality

for synthesis purposes. The work of Hourdin, Charbonneau and Moussa (1997) is foregrounded here, as it has influenced the choice of timbre space for the system proposed in this thesis. The second category takes in a number of studies which have applied techniques drawn from artificial intelligence: in particular, knowledge based systems and, most importantly, computational algorithms based on evolutionary mechanisms occurring in nature. In this section, particular focus will be on the work of McDermott, Griffith and O'Neill (2007). While much of their work was published only in the late stages of preparation of this thesis, it is useful to compare their approach with that taken here.

The chapter begins, however, by revisiting the issue of the visual representation of sound, and reviews two novel interfaces based on posited synaesthetic correspondences between sound, colour and texture.

4.2. Graphical user interfaces for synthesis

Some of the problems which face designers of intuitive GUIs for the specification of sound have already been reviewed in chapter two of this thesis. As was noted, visual representations of sound in the time and frequency domains are not helpful as vehicles for the specification of new timbres; a higher level of abstraction is required which bridges the gulf between the perceptual and acoustical attributes of sound. This level of abstraction can be seen, for example, in *TimbrePainter* which provides a painting interface for additive synthesis (Bylstra and Katayose, 2005). Similar in many respects to *Metasynth* (reviewed in chapter two), it borrows techniques from standard graphics packages to provide the user with the means to modify a two-dimensional representation of a sound where the horizontal axis is time and the vertical axis is the harmonic number. The amplitude of a given harmonic at any moment in time is mapped to brightness.

However, as the authors note, it is difficult to create sounds with any degree of complexity (such as a piano note), not least because such a task presupposes that the user knows what the overall spectrotemporal envelope is.

4.2.1. Synaesthetic approaches

Neither of these two GUIs provide a useful cognitive link between the perception of a sound and its visual representation. *Sound Mosaics* is a prototype GUI which attempts to address this problem by hypothesizing a synaesthetic link between the attributes of sound and those of colour and texture (Giannakis, 2001).

To support this work, a number of psychophysical experiments were conducted by Giannakis, which indicated perceptual links between (aural) *pitch* and (visual) *brightness* and (aural) *loudness* and (visual) *saturation*, such that a two dimensional space could be constructed; and, more interestingly for our purposes, links between visual texture and sonic attributes, from which a three dimensional space could be obtained. Firstly, *sharpness* (defined here as the fundamental frequency for pure tones, and spectral centroid frequency for complex tones) was found to correlate perceptually with *texture contrast*. Secondly, *dissonance*, using Sethares' definition of dissonance as being the sum of all the dissonances between the partials of a complex waveform (Sethares, 2005), correlated inversely with the degree of *repetitiveness* in the texture. Finally, a correlation was found between *compactness* and *texture coarseness* and *granularity*. The variable for sounds along the *compactness* axis was, in effect, harmonicity; sounds were constructed using noise bands centred around six harmonics, whose bandwidth varied along the scale. Importantly, interface prototypes which incorporated this mapping were tested for usability in comparison with a frequency domain representation, in this

case *Metasynth*; it was concluded that the *Sound Mosaics* interface was more comprehensible and intuitive.

Another system (Schatter, Züger *et al.*, 2005) was proposed which similarly sought to exploit synaesthetic links between colour, texture and sound in the design of a GUI for synthesis. The subtractive synthesis engine was driven by twenty-three independent parameters which were mapped to the parameters involved in the modification and manipulation of a graphical 3D object – material (texture), colour, height, width etc. Recognising that such associations are highly subjective and personal, the authors made use of fuzzy logic and genetic algorithms (defined and discussed later) in a 'personalization' task, in which the user made explicit associations between each one of twelve visual metaphors presented and a particular sound. Schatter *et al* reported mixed results, noting that 'the effects of adjusting the metaphors *Width* and *Bulb* (another GUI parameter) are not intuitively understood by the users.'

4.3. Timbre space

4.3.1. Criteria for synthesis

We turn now to reconsider the use of timbre space, discussed in the previous chapter, but this time as a vehicle for sound synthesis. Bearing in mind the discussion in chapter two and three, we can at this point summarise and propose a set of criteria for an ideal *n*-dimensional attribute space which functions usefully as a tool for the search strategy described later in the thesis:

1. It should have good coverage – that is to say, it should be large enough to encompass a wide and musically useful variety of sounds.

2. It should have sufficient resolution and precision.
3. It should provide a description of, or a mapping to a sound sufficiently complete to facilitate its re-synthesis.
4. The axes should be orthogonal – a change to one parameter should not, of itself, cause a change to any other.
5. It should reflect psychoacoustic reality. The perceived timbral difference of two sounds in the space should be broadly proportional to the Euclidean distance between them.
6. It should have predictive power. A sound **C** which is placed between two sounds **A** and **B** should be perceived as a hybrid of those sounds.

Recalling that timbre space is a coordinate space whose axes represent acoustical attributes, chapter two identified two types of timbre space:

- *attribute space*, whose orthogonal dimensions each represent a quantifiable acoustical attribute vector, where the relative distances between objects (sounds) in the space may not reflect or correspond to perceptual differences, and

- *perceptual space*, constructed using (for example) MDS analysis, where each dimension may or may not correspond to a single attribute vector, and where the relative distances between objects in the space reflect and are derived from similarity-dissimilarity judgments.

To this, we must add a third type of space – *synthesis parameter space*, each of whose dimensions represent a single orthogonal synthesis parameter – carrier frequency, cut off frequency etc.

The extent to which points in an attribute space map to points in the parameter space varies according to the chosen synthesis method. Where the parameters of the chosen synthesis method operate at a relatively low level (for example, additive synthesis), the mapping is likely (although not guaranteed) to be straightforward. On the other hand, synthesis methods such as FM, as we have seen in chapter two, do not map in any very linear manner to timbre spaces, and may take the form of lookup tables, where each point in the attribute space is associated with a set of parameter values. While efficient, there is a high cost in storage requirements, particularly for timbre spaces of high dimensionality (Vertegaal and Bonis, 1994). The number of dimensions necessary to fully represent timbral attributes presents computational problems.

Some studies have sought to address this by proposing data reduction solutions. Other researchers have sought to bridge the gap between attribute/perceptual space and parameter space by employing techniques drawn from artificial intelligence. These approaches will be examined in sections 4.3.2. and 4.4 of this chapter.

4.3.2. Data reduction approaches

The multidimensional nature of timbre presents difficulties for designers of control interfaces for sound design; one solution is to use data reduction techniques to create more usable timbre spaces having fewer dimensions. Data reduction of several timbre spaces has been studied, using, for example, principal components analysis (PCA) and factorial analysis of correspondance (FAC) techniques.

4.3.2.1. Principal component analysis

PCA is an analysis technique first described by Hotelling (1933) for identifying redundancies in a multivariate data set (of which timbre space is an example), and reorganising the data so that such redundancies are excluded. It is often the case that two or more variables in a dataset are highly correlated. Where this is the case, a group of such variables can be replaced by a single new variable or *principal component*. The component that accounts for the greatest degree of variance in the dataset is the first principal component; the second principal component, which is orthogonal to the first, accounts for the second greatest degree of variance, and so on. The technique has much in common with MDS – in fact, the results of PCA can be seen as an optimal scaling of those from MDS (Cox and Cox, 2001). However, for our purposes, the crucial difference is that MDS, arising from similarity/dissimilarity judgments between pairs of objects, tries to preserve those pairwise distances in the space.

Sandell and Martens (1995) applied PCA techniques to three musical instrument tones (cello, clarinet and trombone, playing different pitches), downsampled from 44.1 kHz to 22.05 kHz, and analyzed using a phase vocoder. From this data, thirty-one stimuli were reconstructed for each instrument, such that the first was built from just one PC, the second from two and so on. The aim of the study was to establish the number of principal components needed to reconstruct the tones, such that they were a) indistinguishable from the originals, or b) subtly differed from the originals, in such a way that the difference would not be noticed by anybody not alerted to the possibility of difference. The number of principal components (averaged over the results from the three subjects who took part) needed to obtain resyntheses[10] which were indistinguishable from the original was about twenty-two for the cello, fourteen for the clarinet and ten for

[10] Inferred from the graph on page 1022 in Sandell *et al.* Actual figures are not given.

the trombone. In order to attain the second (and lower) threshold of quality, however, the number of PCs was just over ten for the cello, just under ten for the clarinet and about five for the trombone.

PCA data reduction was also used by Nicol (2005) to construct a timbre space. In this study, a distinction was made between *human timbre spaces* and *automated timbre spaces*. Human timbre spaces were defined by Nicol as 'those generated from experiments on human perception', whereas automated timbre spaces are 'generated mathematically by analysing patterns in the sound'. These two types of space seem to correspond to what has been defined and described in chapter three as *perceptual* and *attribute spaces* respectively. The automated (attribute) space used in this study was a Time-Frequency Representation (TFR), generated by analysis using using CQT (Constant Q Transform) and STFT (Short Time Fourier Transform) of real instrumental sounds, such that each frequency is a dimension, and each time step represents a point; a sound is thus represented by a path in this space. Twenty-seven sounds were chosen which closely matched those used by Grey (1977) and Hourdin *et al* (1997). Both CQT and STFT divide the input signal into successive windowed time frames which are then analysed; CQT, however, produces an analysis where the frequency scale is logarithmic, and thus more consistent with human hearing. A PCA data reduction was then generated from the TFR space. The synthesis parameter space was represented as a 'mesh', or cloud of points within the PCA space, such that each point was tagged with a set of FM parameters. Mapping between the two spaces is then done by *proximity detection*, in which the object in the mesh which is closest to the point in the timbre space is used.

However, PCA does not necessarily offer the best way of approximating time-varying spectra. Both PCA and genetic algorithms (discussed later) were

explored in a study to determine a set of basic spectra which would best approximate, for synthesis purposes, the time varying spectra of an original sound (Horner, Beauchamp *et al.*, 1993). Three instruments were analysed; a trumpet, a guitar and a tenor voice, and it was concluded that genetic algorithms were better suited to this task.

4.3.2.2. Multidimensional scaling

MDS has already been discussed in the context of timbre studies in the previous chapter. However, Hourdin, Charbonneau and Moussa (1997) demonstrated that an MDS space of reduced dimensionality, created, not from psychoacoustic listening tests, but from physical descriptions of the sound had potential use as a synthesis space. Because the design of the timbre space used in chapter seven of this thesis draws extensively on this work, the aims, methodology and findings of these two key papers are discussed here.

A set of forty orchestral instrument tones was compiled, partly based on that used by Grey (1977), but also including a number of others, such as a marimba, tubular bells, and a harp. These tones were subjected to *heterodyne filtering* using the *Csound* programming language. Heterodyne filtering resolves periodic or quasi-periodic signals into component harmonics, given an initial known fundamental frequency (Freedman, 1965; Freedman, 1967; Beauchamp, 1969; Moorer, 1973; Beauchamp, 1975; Moorer, 1975). The multiplication of the input waveform by a sine and cosine function at the fundamental frequency and at integer multiples of that frequency and the summing of the results over a short time period yields amplitude and phase data for each harmonic. This can be summarised more precisely as:

$$A_k(n) = \sqrt{a_k^2(n) + b_k^2(n)}\cdot$$ is the estimated amplitude value of the *k*th harmonic at time *n*T,

$$F_k(n) = \frac{1}{2\pi}\frac{\delta\theta_k(n)}{dt}$$ is the estimated frequency value of the *k*th harmonic at time *n*T,

$$\theta_k(n) = \text{atan}\frac{a_k(n)}{b_k(n)}$$ is the estimated phase value of the *k*th harmonic at time *n*T

and

$$a_k(n) = \sum_{r=n}^{n+N-1} x(r)\sin(rk\omega_0 T)$$

$$b_k(n) = \sum_{r=n}^{n+N-1} x(r)\cos(rk\omega_0 T)$$

where **x(r)** is the input waveform , **T** is the sample period, $\mathbf{w_0}$ is the radian frequency of analysis and **N** is the number of samples in one cycle of the input waveform, rounded to the nearest integer (Hourdin, Charbonneau *et al.*, 1997).

While heterodyne filtering is a useful tool for analysis/resynthesis, there are limits to the method; it operates effectively only on steady state tones where the pitch does not change by more than a quarter tone, and where the attack time is not less than 50 milliseconds (Moorer, 1973) ; outside those restrictions, it becomes increasingly inaccurate, although a tracking version was developed that could follow pitch variations (Beauchamp, 1981). For the purposes of the study presented by Hourdin *et al*, however, the technique provided an adequate means of representing each of the forty tones described above. Because this same technique was used in the empirical work discussed in chapter seven, we describe it in some detail here.

Hetro, the Csound implementation of heterodyne filtering, generates, for each spectral component (harmonic) of the sound (the exact number can be specified) two vectors of values which describe the frequency and amplitude fluctuations of that harmonic with respect to time. Thus, for a pitched sound whose fundamental was 250 Hz the following function call

```
hetro -f 250 –h 80 audiofile.aif hetrofile
```

would generate a matrix of eighty (40 x 2) columns from the input sound file *audiofile.aif*, storing it in a file called *hetrofile*. Each row of this output file is a spectral snapshot of the sound; the file thus contains additive synthesis control data for the reconstruction of the sound(s), which can be done using the Csound function *adsyn*.

However, the purpose of this study was to reduce the dimensionality of this dataset before resynthesis. The eighty column matrix generated was analysed using MDS [11].

As discussed in the previous chapter, the input to an MDS analysis is typically a matrix of 'proximities' (which may represent similarity/dissimilarity judgments, for example) between a set of objects. (In this case, the 'proximity' matrix was generated from the heterodyne output, before being analysed.) The output is a geometric configuration of points, each representing a single object in the set, such that their disposition in the space approximates their proximity relationships. Each dimension of the space is called a 'factor'; the first one captures the maximum variance of the data, the second captures the second highest amount of variance and so on. Hourdin *et al* found that 85 per cent of the variance in the

[11] In this study the program used was ANCORR, a package which implements a factorial analysis of correspondences (FAC) algorithm.

heterodyne data could be contained within six factors; that is to say, a six dimensional solution could be generated in which the relative distance relationships between the objects (each corresponding to a row or spectral snapshot in the original heterodyne data) reflected their distances in the eighty dimensional space.

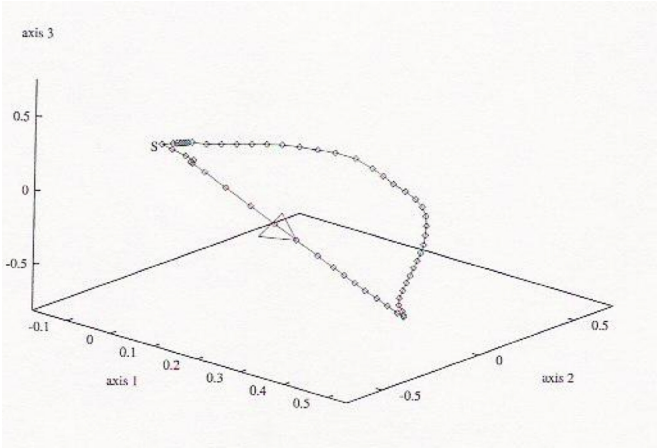


Figure 4.1: Flute tone trajectory (from Hourdin *et al* 1997a)

Figure 4.1 shows the trajectory of a flute, projected onto the three factors which represent the most variance in the MDS space. Each discrete point on the curve represents a single spectral snapshot in the heterodyne data matrix.

It is crucial that the reduced space generated by the MDS process is stable – i.e. that the addition of a new instrumental sound to the analysis, or the removal of an existing one does not significantly affect the shape of the space or the distance relationships between the other sounds in the space. In order to verify this, a new MDS analysis was generated, omitting twelve sounds that were present in the original one. It was found that the two spaces did not significantly differ.

This reduced space provided the basis for the synthesis space proposed in their second paper (Hourdin, Charbonneau *et al.*, 1997). The software that they used allowed the reconstruction of the original heterodyne data, from which the sounds can be synthesised by additive synthesis. Clearly, the original heterodyne

data can only be recovered with 100% accuracy if the number of factors is equal to the number of dimensions in the original input data. Repeatedly rebuilding the data using an increasing number of factors resulted in a corresponding increase in accuracy, with a significant improvement in quality when a seven factor analysis was used. Hourdin *et al* found that the first two axes of the space corresponded to the first two axes generated in Grey's MDS study (1977) (discussed in chapter three) , suggesting that the space has perceptual validity.

Equally importantly, the reduced space permitted the synthesis of new sounds by creating a curve in the space that was a linear interpolation between two existing curves. Hourdin *et al* demonstrated this by plotting two curves that lay between the curves for the tenor trombone and the *martelé* cello. Again, although no formal listening tests were apparently conducted to verify this, the authors reported that the new sound seemed audibly to be a hybrid of these two instruments.

The whole process of analysis and resynthesis is summarised in figure 4.2. and will be revisited in chapter seven.

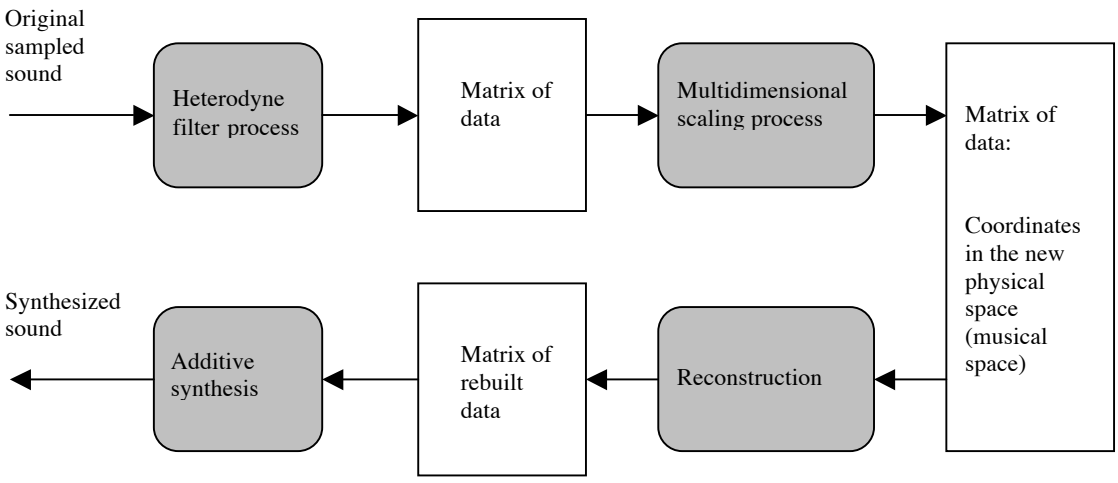


Figure 4.2: Analysis-resynthesis process – from Hourdin *et al* (1997b)

4.4. Artificial intelligence methods.

4.4.1. Introduction

The other approach to bridging the gap between timbre space and parameter space draws on artificial intelligence technology.

Artificial intelligence is the research field concerned with the design and study of systems which exhibit intelligent behaviour (so called 'strong' AI) or which emulate some facet of intelligent behaviour – reasoning, learning, perceiving, communicating or planning ('weak' AI). Research in this area rests on the proposition that intelligence, or aspects of it, can, in principle, be simulated by a machine; AI methods are typically employed in application domains requiring deduction, planning, problem solving, learning, perception and creativity.

AI approaches to the specification of sound have taken a number of forms, which for the purposes of this thesis, can be broadly grouped into two categories of interest. These are

- knowledge based systems (KBS); such systems are based on explicitly encoded rules and heuristics which relate to synthesis expertise, or to the mapping of specific acoustic attributes with the adjectives and adverbs used to describe sound.
- evolutionary search strategies such as genetic algorithms (GAs), cellular automata and artificial immune systems.

In the following sections, we define and discuss these two categories and consider their application in the design of systems for sound synthesis.

4.4.2. Knowledge-based systems (KBS)

Chapter two reviewed research whose aim was to determine the correspondences between adjectives and adverbs typically used to describe sound and specific sonic attributes. This section revisits this subject, but this time in order to examine synthesis systems which seek to bridge the semantic gap in the user/system interface between language and synthesis parameters. Many of these systems encode rules and heuristics for synthesis in a knowledge base.

The difficulties of precisely defining the relationship between a given point in timbre space and a point in a parameter space have prompted some researchers to frame the problem as one of knowledge representation. KBS systems typically make use of heuristics such as 'bright sounds have significant energy in the upper regions of the frequency spectrum', 'a whole number modulator/carrier frequency relationship will generate a harmonic sound' etc.

A knowledge based system is typically one which encodes facts about a particular domain, together with, for example, *if-then* rules which can be used by an inference engine for problem solving and diagnosis. The principles on which a knowledge based system for sound design might be built have been explored by Miranda (1995). Starting out from the premise that design is "an explicitly knowledge-based kind of intelligent behavior", Miranda lists a number of desirable characteristics of an Intelligent System for Sound Design (ISSD). It should offer a user interface which facilitates communication using a 'intuitive, perceptually oriented vocabulary', rather than one which permits only low level, and essentially numerical directives. The system should be capable of being configured according to the user's own terminology and sound world; should offer intelligent support in the design task; and, lastly, should exhibit the most

important characteristic of any intelligent system, the ability to learn, and to update its knowledge base. The ISSD system will be reviewed later in this section.

Systems for sound specification which have been built using KBS methodologies tend to fall into one or more of the following categories; it should be noted that this is not an exhaustive list, nor are the categories mutually exclusive. Such systems encapsulate and encode

- synthesis expertise (e.g., FM synthesis) – the correct parameter settings for a given sound (Ashley, 1986; Vertegaal and Bonis, 1994; Miranda, 1995; Rolland and Pachet, 1996; Miranda, 1998)

- knowledge of how a sound may be transformed (Ethington and Punch, 1994; Rolland and Pachet, 1996)

- knowledge of how a sound may be blended from two 'parent' sounds (Martins, Pereira *et al.*, 2004)

The following sections examine a number of representative systems above, these three broad categories.

4.4.2.1. Synthesis expertise

One of the earliest ventures into this area was *TD*, a KBS system which drew on a knowledge base linking verbal labels with the parameters of FM synthesis (Ashley, 1986). Noting the problems of using FM - the lack of a perceptually clear link between parameter values and the audible attributes of sound, and the fact that it does not present a uniformly shaped search space - *TD*

was built using a production system architecture, consisting (amongst other components) of a *learning system* and a set of *timbre frames.* A timbre frame contained knowledge and heuristics relating to some verbal descriptor (e.g., *bright*, *brassy* or *bite*), and the synthesis parameters and parameter values which could be used to generate a sound having this attribute. These could be coded by the programmer or, importantly, could be acquired from the user interactively, using the learning system.

The learning system was based on inductive acquisition of heuristic knowledge, achieved through a process of hypothesis formation and testing. Two sounds were presented, which varied in the value of one FM parameter; the user was then asked to describe the difference (if any) in timbre between the two sounds. The system then attempts to generalise from this information in order to apply it in different situations; this is then used to build rules and parameter value minima and maxima for a timbre frame. This system is of particular interest in that, unlike a number of later systems, which made use of information previously acquired from listening tests, it iteratively built up a rule base tailored to the individual user.

Rolland and Pachet (1996) more recently proposed a system for capturing expertise in FM programming using a model of knowledge representation based, not on the attributes of sound structures themselves, but on the transformations that can be applied to those sound structures. They observe that ‘many famous patches were designed by people who understood only a limited fraction of the underlying FM theory.’ In this model, the knowledge base consists of rules on, for example, the transformation procedure to be followed to make a sound 'brighter' or 'warmer'. This, in turn, necessitates classification of a sound according to the transformations that it can undergo (brassy-able, capable of being made brassy) rather than its own inherent properties (‘brassy’). In this way, a hierarchical

network of sound-types can be built up, in which a connection between any two components defines the transformation that changes one into the other.

Another approach to the capturing of synthesis expertise is the Intuitive Sound Editing Environment (ISEE) (Vertegaal and Bonis, 1994). It proposes a hierarchical structure of timbre spaces (here referred to as 'instrument spaces'), based on a taxonomy of musical instruments. Thus, at the 'root' level, the classification is according to whether the sound is *sustaining* or *decaying;* classification at the next level of instrument space is on the basis of *harmonicity* or *inharmonicity* and so on. The 'leaf' instrument spaces of this tree structure are specific instrument types - clarinet, vibraphone etc . Each instrument space is a timbre space implementation whose axes are *overtones* (basic harmonic content), *brightness* (spectral energy distribution), *articulation* (spectral transient and persistent noise behaviour) and *envelope* (temporal envelope speed). The user is able to 'zoom in' to an instrument space which is solely occupied by one instrument (marimba, trumpet), or alternatively 'zoom out' to a instrument space which encompasses the characteristics of an instrument type (plucked, bowed, sustaining etc). Underpinning this is an 'interpreter' which translates parameter values into MIDI System Exclusive data tailored towards the particular synthesizer; thus, a change in timbre requiring numerous parameter changes can be effected by relocating the sound within the instrument space hierarchy.

The ISSD system proposed by Miranda (1995) and which draws on Schaeffer (1966), Chion(1983) and Slawson (1985, 1987), is built on an Abstract Sound Schema (ASS); this schema represents a sound in terms of its attributes (*brightness*, *openness*, *compactness*, *acuteness* etc) and holds information on how these attributes are mapped to synthesis parameters. Using inductive learning techniques, the system is able to 'learn' new sound events from incomplete data, or from new data input by the user. Because the system implements notions of

'class', 'instantiation' and 'inheritance' it can in some respects be seen as an object orientated programming methodology for sound. These ideas were implemented in a system called ARTIST (Miranda, 1998) which allowed the composition of sounds conceived in descriptive rather than quantitative terms.

4.4.2.2. Transforming

SeaWave (Ethington and Punch, 1994) is structured around a family of timbres, each of which can be varied to produce new timbres by applying *transformations*. The distinctive feature of this approach is that timbre control is provided by presenting the user with the means to change some existing timbre. Unlike the two previously described systems, the mapping of verbal descriptors to synthesis parameters was derived through listening experiments (although it is not apparent from the paper whether these were structured listening tests involving a number of subjects).

An initial list of 124 adjectives was categorised into 'similarity' classes, according to the particular phase in the temporal evolution of the sound which they described. Thus, the *attack* class contained terms like *blown*, *bowed*, *hammered*, *plucked* etc., whereas terms such as *percussive*, *sustained*, *damped* were placed in the *cutoff* class, defined here as the manner in which 'the sound becomes inaudible'. The seven remaining classes each contained terms relating to some aspect of *presence*, defined here as 'the quality of the sound while it is sustained'. *Presence Class A*, for example, contained terms which were to a greater or lesser extent synonymous, like *airy*, *breathy*, *windy*, *blowing* etc.; similarly, *Presence Class B* was characterised by adjectives such as *warm*, *ringing*, *clear* etc.

A set of tones whose acoustical attributes varied widely was then created; each tone in this set was then used as the basis of a set of tones in which one attribute (e.g., harmonic distribution, weight of partials etc) was systematically varied, the variable in each case being referred to as an *operator*. Each series was auditioned, the purpose being to determine the extent to which the adjectives correlated with operators. (This approach notably differs from other studies, such as those described in the previous chapter, in that the emphasis is on mapping an adjective to a transformation in some aspect of the sound, whereas other studies have sought to establish invariances in sounds which are described as similar.) It was found that some adjectives were not used at all to describe a given operator and, conversely, other adjectives were associated with more than one.

The interface offers the user a 'parent' timbre, together with the means to select a particular adjective from each class (*attack*, *presence* and *cutoff*) and then to apply transformation to the sound ('*slightly more plucked'*, '*much less resonant'*, '*more damped'* etc) to generate new 'child' timbres.

4.4.2.3. Blending

Divago (Martins, Pereira *et al.*, 2004) was a system inspired by ARTIST (described above) which made use of both knowledge representation techniques and genetic algorithms (discussed later in this chapter). In this package, a pair of sound descriptions could be *blended* to form a new one which inherited characteristics from its 'parents' but may also contain new ones inferred from (for example) a rule base. It consisted of a knowledge base containing *semantic networks* for sounds. (A semantic network is a set of concepts linked by semantic relationships – e.g., *dog* and *animal* are linked by the semantic relationship *is a*.) Thus, a sound could be defined in the same way as in ARTIST – its *pitch* could be *high*, its *duration*

could be *long*, and its *timbre bright* and so on. The system can then blend this with another sound, some of whose attributes may be different, by proposing a number of possible candidates whose fitness is then evaluated by a genetic algorithm according to a number of constraints based on those proposed in a paper on optimality in conceptual blending (Pereira and Cardoso, 2003) .

4.4.2.4. Other approaches

More recent approaches have mapped descriptors to synthesis parameters using other methods. A system proposed by (Howard, Disley *et al.*, 2007), again making use of an additive synthesis engine, features eight sliders, each of which maps a particular descriptor to an MDS dimension. Using a set of instrumental sounds from a sample library (viola bowed, viola pizzicato, electric guitar, tenor trombone, bach trumpet, trumpet harmon, flute vibrato, alto saxophone, oboe, hamburg steinway, tubular bells, and xylophone), a set of listening tests were conducted (similar to those described in the previous chapter) (Disley, Howard *et al.*, 2006) which established a list of adjectives which subjects felt they could use with most confidence for describing the stimuli sounds or where there was the highest degree of agreement across subjects. From this list, the words *bright*, *clear*, *warm*, *thin*, *harsh*, *dull*, *percussive* and *gentle* were selected as controls for the synthesizer. A further listening test was used to map these terms to two dimensions of an MDS solution, in which the term *percussive* was associated with one dimension and the remaining terms with the other dimension. The synthesizer itself was built in PD, a freeware GUI-based graphical programming environment similar to Max/MSP.

A broadly similar list of adjectives (*bright*, *warm*, *harsh*, *hit*, *plucked*, *constant*, *thick*, *metallic* and *woody*) has been used in a system which, again, provides a set of

sliders corresponding to these adjectives, but in which the mapping is achieved by training a neural network (Gounaropoulos and Johnson, 2006). The system uses this data to classify an input sound and sets the values of the sliders accordingly. The user then adjusts the values of the sliders, and a search is conducted. This information is then used to search through the synthesis parameter space. The authors have compared the operation of two search methods; the first makes use of a genetic algorithm, while the second one involves feeding the new values back into the neural network. At the time of writing, the neural network method seemed to produce better results; the authors note that while GAs return a single fitness value, the neural network method returns a separate error value for each parameter, enabling faster convergence.

A recent proposal for timbral synthesis using language is one which is based on the eight timbral descriptors in the MPEG-7 standard (Mintz, 2007). These descriptors derive from much of the psychoacoustical work reviewed in the previous chapter; for harmonic sounds (that is to say, those with harmonic spectra), these are log-attack time (*lat*), harmonic spectral centroid (*hsc*), harmonic spectral standard deviation (*hsstd*), harmonic spectral variation (*hsv*), and harmonic spectral deviation (*hsd*). For percussive sounds, the descriptors are (again) log-attack time (*lat*), temporal centroid (*tc*) and spectral centroid (*sc*). The user interface provides a number of sliders for values like *bite*, *brightness, warmth* etc which are mapped to the MPEG-7 descriptors for synthesis.

4.4.2.5. Limits

The problems and limitations of mapping verbal descriptors to the measurable features of sound have already been examined in section 3.4.4.4 of the previous chapter on timbre perception, and are recapitulated here:

- There is a complex and non-linear relationship between a timbre space and a verbal space.
- There are questions of the cross-cultural validity and common understanding of descriptors.
- The choice of descriptors for a given sound is likely to vary according to listener constituency.
- Apparently similar semantic scales may not actually be regarded by listeners as similar.

These issues apply with equal force when we consider synthesis. In addition, the context in which the computer musician is working may influence the choice of terms used; for example, an instruction to 'play louder' has a different meaning in the context of music by Morton Feldman (to take an extreme example) and Richard Strauss (Ashley, 1986).

4.4.3. Evolutionary search algorithms

Having considered a number of representative knowledge based systems designed to bridge the user/system language gap, we turn to another class of systems which consider the problem of arriving at a desired sound in a given synthesis space as one of search, and the process of doing this to be one of making incremental evolutionary changes to the properties of a candidate sound.

In general, a search algorithm is a computational procedure designed to return a solution to a given problem from a set of all possible solutions, called the search space. The search space may be tree-structured or graph-structured, such that each node of the tree represents an object in the search space. *Depth-first* and

breadth-first algorithms search such spaces by visiting and evaluating each node successively; such algorithms are *blind*, *naïve* or *uninformed*, and are expensive in terms of the time required to satisfy the goal. By contrast, *heuristic* searches makes use of information about the structure of the search space to direct the choice of which node to visit next. As each node is visited, a heuristic function (a rule) is called which returns a value reflecting how promising an exploration starting from this node would be. Typically, this is expressed as a *cost* – the lower the value, the greater the probability that the goal will be satisfied. *Best-first* search orders the search in order of cost.

A number of search algorithms draw on evolutionary mechanisms found in nature. First developed by Bienert, Rechenberg, and Schwefel in the 1960s for the optimisation of body shapes for minimal drag in wind tunnels (Bäck, 1996; Beyer and Schwefel, 2002), evolution strategies (ES) provide a method of optimisation in which a random set of candidate 'parent' solutions are mutated to generate new 'child' solutions. Two versions of this strategy have been developed since then, known as $(\mu/\rho+\lambda)$ **–ES** and $(\mu/\rho, \lambda)$**-ES**, where μ is the number of parents, $\rho \leq \mu$ is the number of parents selected for mutation and λ is the number of offspring. The first version selects the best μ individuals from just the child population, whereas the second selects from both child and parent populations. In both cases, the selection is on the basis of some predefined fitness function, and the process is repeated.

4.4.3.1. Genetic algorithms

Genetic algorithms (GAs) have much in common with the search strategies describe above, and can be seen as a special category of ES. They provide a means of arriving at an optimal solution within a search space (Holland, 1975), by

encoding (usually, but not always, in binary form) a population of possible solutions, evaluating each solution using a problem-specific *fitness function*, allowing the 'best' solutions to breed new solutions, and iteratively re-evaluating them. By a process analogous to naturally occurring evolutionary mechanisms – mutation and crossover of 'genetic' material, successive populations are bred whose fitness for purpose is improved on each iteration.

One particular version of the algorithm is *binary tournament selection*; this consists of a *selection* phase in which members of the old population are randomly paired up, and the 'best' one of each pairing goes forward to the new population. This process occurs twice, thus ensuring that the new population has the same number of individuals as the old one. In order to generate greater diversity in the new population (and therefore possible solutions), a number of individuals in the new population 'mate' with others through *crossover*, in which the two individuals swap a predefined number of bits. This step can be disruptive, in that the new individuals may well be poor candidates ; however, it also allows the possibility of interesting new combinations. Finally, there is a *mutation* phase, in which each bit may be randomly changed, based on a probability which is related to the reciprocal of the bit length.

Genetic algorithms in particular have been applied to a number of problems related to music performance and composition. A system by Horner and Goldberg (1991) generated a thematic bridging between a initial and a final musical pattern, both prespecified, in a manner characteristic of minimalist music. A number of transformation operations – deletion or mutation of a note, rotation of the pattern etc - were defined; the GA then arrived at an optimised sequence of these operations through which the final pattern arrived at most closely resembled the specified one. *Genjam* (Biles, 1994) was a system designed to simulate the process

of a jazz musician learning to improvise, where the fitness function, supplied by a human 'mentor', was used to generate new musical phrases. *Vox Populi* (Manzolli, Moroni *et al.*, 1999; Moroni, Manzolli *et al.*, 2000) used GAs for algorithmic composition of chord sequences, in which the degree of internal consonance, the consonance between the notes of the chord and the tonal centre and the extent to which the individual notes fitted within the standard SATB[12] voice ranges were the fitness criteria.

GAs have also been employed in sound synthesis. Dahlstedt notes that evolutionary methods can be seen not simply and solely as a means of optimising the parameter values of a particular synthesis method; they also offer a useful means of exploring the parameter space (Dahlstedt, 2007).

The problems of mapping derivation of complex spectra from carrier index and carrier/modulator frequency ratio parameters - the essence of frequency modulation (FM) synthesis - can be seen as a one-way function, in that it is very difficult to reverse engineer the required parameter values from a given spectrum. This is a search problem that lends itself to solutions using either evolutionary strategies or genetic algorithms.

A 1993 paper presented a GA for finding optimized parameters for FM synthesis of a sound already analysed using short time Fourier analysis (Horner, Beauchamp *et al.*, 1993). The parameter values to be found were the frequencies and amplitudes of the single modulator and multiple parallel carriers between the target spectrum and spectra generated from candidate solutions, where the ratio of carrier frequency to modulator frequency was an integer (thus producing only harmonic tones). This work was extended by Mitchell and Pipe (2005), who

[12] Soprano, Alto, Tenor, Bass

demonstrated an effective evolutionary strategy for determining FM parameters where the carrier/frequency ratio is not necessarily an integer, and therefore makes a wider range of sounds accessible. The synthesis model was single carrier/modulator, and the strategy based on an ES working on a (5/5, 25) basis (i.e. 5 parents and 25 offspring). This differed from the work of Horner, Beauchamp *et al* in that the algorithm was user driven - the fitness of the child solutions was evaluated aurally. Interactive evolutionary strategies, of which this is an example, will be considered later in this chapter.

FM is not the only synthesis method to which GAs have been applied. Another study made use of a genetic algorithm to determine the group synthesis parameters required to reconstruct a previously analysed sound (Cheung and Horner, 1996). *ESSynth* (Manzolli, Maia Jr *et al*., 2001) was based on an interesting approach, in that the waveform vector itself was designated as the genotype, rather than some binary representation of it . In this system, crossover was done by exchanging waveform segments between individuals in the population, mutation by 'waveshaping' the waveform by another random waveform by an amount determined by a random mutation coefficient. A more recent version of it was evaluated both objectively and subjectively by Caetano, Manzolli *et al* (2005), although it is unclear exactly how the subjective evaluation was conducted.

In the case of *ESSynth*, the fitness function for matching candidate sounds against a target was based on the Euclidean distance between them in the vector space that they occupied. Clearly, however, fitness functions can be based on a number of different criteria; for example, a *pointwise metric* (the pointwise difference between the two functions), a *DFT metric* (the difference between the normalised DFTs of the two functions), *perceptual metrics* (based on calculated differences in harmonicity, centroid, and attack time), or on a weighted composite

of all of these. McDermott, Griffith *et al* (2005) performed a comparative analysis of a GAs operating on waveforms with one, four, eight, sixteen and fifty partials whose fitness functions were based on each of the above, as well as on a combination of them, and concluded that a composite fitness function drives evolution more successfully than fitness functions which use only one of these criteria.

A genetic algorithm was applied to a physical modelling synthesis problem by Riionheimo and Välimäki (2003). In order to simulate and synthesize accurately the vibrations of a plucked string, it is necessary to model both its horizontal and its vertical motion – that is to say, its motion parallel and perpendicular to a soundboard. This can be achieved by the use of two slightly mistuned string models; however, optimal values for the nine interacting parameters required for this synthesis are difficult to determine. Riionheimo *et al* presented a system for automatic parameter extraction from recorded string tones using a perceptual fitness function which made use of an auditory model of human hearing.

4.4.3.2. Genetic programming

Genetic programming or GP (Koza, 1992) can be seen as a special case of GA, in which hierarchical tree structures which may variously represent computer programs or mathematical functions are subject to crossover and mutation operations. Crossover, however, takes place by the interchanging of randomly chosen branches of the two parent trees to create two new offspring; mutation, by the random selection of subtree, replacing it by another randomly generated subtree. Because they operate on structures whose elements are functional components, GP techniques have been used to evolve electronic circuits as well as computer programs. Garcia (2001) applied a GP to the design of synthesis

topologies consisting of oscillators, filters etc, where, again, individuals in the population were evaluated on the extent to which the sound generated by the topology matched a previously specified model.

4.4.3.3. Interactive evolutionary strategies

Of greater interest to the present discussion are GAs in which the fitness function is not the minimisation of the difference between candidate solutions and some pre-existing model, but instead is determined by the user using any criterion – subjective, objective, aesthetic – he/she chooses. These are *interactive genetic algorithms* or IGAs, first proposed by Richard Dawkins (Dawkins, 1986; Dawkins, 1988) in the "biomorph' system for exploring evolutionary mechanisms. Since then, they have been successfully applied, both in the musical domain (*Genjam,* mentioned above, was one such system), and to problems of parameter optimisation in synthesis. The drawback is that user-driven evaluation, by its very nature, cannot be easily automated (if at all), because of the difficulty in explicitly defining the criteria (Dahlstedt, 2001).

An example of such a user-driven algorithm was proposed by Takala *et al* (1993). In this system, mathematical functions required to generate a particular waveform were represented as tree structures or 'timbre trees', the nodes of which variously represented numerical constants, variables or functions using sub-trees as arguments. This approach is not unlike that of Garcia (2001) except that the fitness function here is user-driven. In Takala's system, genetic algorithms were used to mutate timbre trees, and their evolution guided by user input.

Typically, the selection of individuals for breeding in, for example, *binary tournament selection* (discussed in section 4.4.3.1), is done by truncation: weaker

individuals are eliminated and only reappear in the population as a result of crossover and/or mutation. However, the application of of interactive GAs to sound synthesis carried out by Johnson (1999) was distinctive in that the new population was generated based on *fitness proportionate selection*, in which the higher the fitness rating given to an individual, the more likely it is to be selected. Because the fact that an element of probability is involved has some bearing on, and relevance to the probability table technique presented in chapters six and seven, this selection method is introduced and discussed here.

The IGA presented in Johnson (1999) is a front end to Csound, and is designed to optimise seven of the parameters used in the *fof* algorithm, used for granular synthesis. Each of the parameters is encoded as a sixteen bit integer, concatenated into a 16 x 7 = 112 bit binary string. The procedure randomly generates a population of nine 'sounds'. The user then assigns a rating $\mathbf{r_s}$ to each individual 'sound' **s** in the population; for each sound string, a probability $\mathbf{p_s} = \mathbf{r_s/t}$ is calculated, where **t** is the sum of all the ratings**.** These probabilities are then used in a *roulette-wheel selection* procedure to choose pairs of parents for the next generation (Goldberg, 1989), illustrated below.

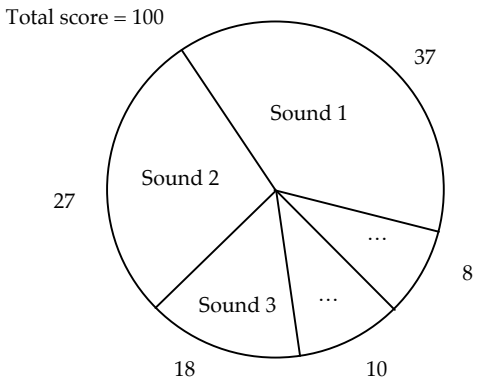


Figure 4.3: 'Roulette wheel' selection (from Johnson (1999)).

In fig 4.3, sound 1 has a 37% probability of being selected to be a parent, sound 2 a 27% probability and so on. This means, of course, that weaker

candidates may be carried forward; the advantage is that it allows the possibility of revisiting parts of the search space previously rejected.

4.4.3.3.1. Mutasynth

Developed by Dahlstedt (2001), *Mutasynth* was a generic IGA-based system for sound generation which was not tethered to any one synthesis method. The initial population of nine could be transformed not only by the standard genetic operators of mutation and crossover, but also through *insemination, morphing* and manual mutation. In the insemination operation (essentially a variation of crossover), genes from one parent were copied unaltered to the new genome; then a number of genes in the new genome were then overwritten by the corresponding genes from the other parent, the exact number being determined by an *insemination amount* value. Another setting specified whether these genes were contiguous or randomly scattered. Morphing was done by linear interpolation at the gene level; a new genome was formed at a random point on a line joining the two parents in the multidimensional parameter space. Finally, through manual mutation, users were able to manually change parameters on the synthesizer; this change was then reflected back into the gene pool, and all relevent genes are modified accordingly.

One important feature of this system was the facility to 'lock' one or more parameters if the user wished to prevent their disruption by further mutations and crossovers; the relevant genes were 'disabled' and simply passed on unchanged to subsequent generations. (This, of course, raises the question whether the naïve user would know which parameter(s) should be locked in order to retain a particular attribute.)

The interface offered to the user is of interest. Sounds could be stored in a gene bank, to be brought back into the breeding process if required; parents could be selected from a previously stored genome, from any sound in the current sound engine or any individual from the current population. Current sounds were visually represented on screen by wiggly lines whose shape was a depiction of the synthesis parameter values, and which served as a mnemonic to the user. A successor to *Mutasynth*, called *Patch Mutator*, has been developed (Dahlstedt, 2007; Dahlstedt, 2009): unlike *Mutasynth,* which was generic and could be linked to any one of a number of synthesis engines, *Patch Mutator* is integrated into the environment of a Nord Modular G2 synthesizer .

4.4.3.3.2. Genophone

Developed by Mandelis (Mandelis, 2001; Mandelis, 2002; Mandelis and Husbands, 2006), *Genophone* is a system for 'the creation of novel sounds and exploration rather than designing sounds that satisfy specific *a priori* criteria'. In addition to providing a user interface for synthesis, it also affords real time control of synthesis parameters using a dataglove (recall the distinction made in chapter two between 'real time' and 'fixed' synthesis controllers. We confine the discussion to the 'fixed synthesis' element of the system).

Like *Mutasynth,* the software can be linked to a number of different synthesis engines; the breeding process generates MIDI System Exclusive messages tailored to the particular synthesis method; in addition, the system provides a degree of control over the direction taken by the evolutionary process by enabling the 'locking' of selected parameters. As with the other systems described above, the GA is interactive; a 'population window' is provided for the selection process, in which individuals (patches) are positioned such that the

preferred ones are at the top. Crossover takes a number of forms, including interpolating crossover where the unlocked parameter values used for overwriting are located between the parents in the parameter space, weighted by their fitness.

4.4.3.3.3. The fitness evaluation 'bottleneck'

Genetic algorithms take many generations to converge on a solution, and one of the drawbacks of the interactive evolutionary approach is that human evaluation of each individual in the population is inevitably slower than in systems where the determination of fitness is automated (Takagi, 2001). The problem can be minimised, to some extent, by ensuring that only a few candidates at any one time are bred for evaluation (Dahlstedt (2001), for example, worked with a population of only nine).

However, an interesting technique for addressing this problem was proposed by McDermott, Griffith *et al* (2007). Much of this work was conducted at the same time as the empirical work presented in this thesis, and, while there are significant differences in aims, objectives, approach and methodology (the work presented here, for example, does not make use of genetic algorithms), it will be useful later to make a broad comparison of the results from this study and from McDermott *et al*. For this reason, we will consider this particular paper in more depth.

The purpose of the research was, firstly, to compare an interactive evolutionary procedure with one that was not evolutionary; and secondly, to propose and test a novel method to alleviate the fitness evaluation bottleneck - that of interactive interpolation or 'sweeping'.

Four distinct GUIs were constructed for the Xsynth-DSSI synthesizer. The first of these (GUI 0, referred to in this discussion as 'Sliders') presented the user with sliders; each slider represented one synthesis parameter (e.g., LFO frequency) and was labelled accordingly. The second one (GUI 1, which we refer to here as the 'IGA' version) was based on that of Johnson (2003), discussed earlier in this chapter. In this version, a population of sounds were made available for evaluation. The user was presented with a panel containing, for each sound, a radio button and a slider. The radio button enabled selection of the sound, and the user was then required to give a rating to it by setting the value of the slider.

The third GUI (GUI 2, which we refer to here as 'Sweep') introduced the interactive interpolation element. A panel containing a single slider was made available to the user. Three sounds, L C and R could be accessed by placing the sliders at the left, centre and right points of the scale respectively. Points in between were 'mixtures' of L and C and of C and R, i.e. characterised by parameter values which were located between those of L and C and of C and R. This allowed the user to manually control an interpolation at the genetic level between pairs of individuals. A selection having been made, this became the C point, and random individuals were then generated for L and R.

The fourth GUI ("Sweep with background evolution") was a variation on the third. In this version, a target waveform was loaded, against which an automatic EC process is performed in the background, while, at the same time, the 'Sweep' interaction described in the previous paragraph was GUI 2 was run in foreground. The best individual found by the background process was then used as L, instead of being randomly generated.

In all versions, two target sounds were used, both obtainable using the Xsynth-DSSI synthesizer (and therefore reachable using any of the four GUIs) . The first one (Target 0) was a xylophone like sound, while the second was a 'synth strings' sound (Target 1). The nature of these two sounds meant that they occupied very different areas of the search space.

The experiment was preceded by a series of listening tests, whose purpose was to establish how good subjects were in discriminating between sounds. Subjects were asked to listen to three sounds A, B and C and say which of B or C was most similar to A. In each triplet, either B or C was identical to A. The results showed a high success rate, and indicated that all subjects were capable of discriminating between different sounds.

The metrics used to assess the GUIs were a) user rating (defined as user satisfaction with the match), b) the time taken (in seconds) and c) attribute distance (i.e. the distance from target sound to achieved sound), expressed as attribute distance (using a set of forty timbral, perceptual and statistical attributes from the sounds), DFT distance, and the distance between the synthesis parameter values.

The overall significant result was that users required less time to converge on the target using the Sweep GUIs, regardless of target, suggesting that this technique provided a better and more effective interaction than the use of parameter sliders. Other results were less clear cut. In general, there were no statistically significant differences in the user rating between GUIs; however, the 'Slider' GUI was rated more highly for the xylophone target (Target 0), whereas both the 'Sweep' GUIs received a better rating when applied to the synth string sound (Target 1). McDermott *et al* attribute this to envelope matching being more

easily achieved using sliders.[13] The paper concluded that, overall, both the 'Sweep' interfaces were at least as good as, and in some ways better than the other interfaces, as judged by user ratings, attribute distances, and time spent on the task.

The approach taken in this study, together with its findings, will be reconsidered in the conclusion of this thesis.

4.5. Conclusions

This chapter has reviewed a number of approaches to the bridging of the gap between task language and core language, between sound as it is apprehended and described and the tools available for its creation. Two main categories have been identified.

The first of these is where researchers have treated the problem as one of dimensionality reduction in order to make the space more tractable. For the purpose of this thesis, the work of Hourdin *et al* (1997) is of particular importance in providing a baseline for a timbre space to be explored by the search algorithm proposed here. Techniques used in the second category are drawn from artificial intelligence: knowledge based systems which encode in a knowledge base synthesis expertise or rules for generating synthesis parameter values from adverbs and adjectives; and evolutionary algorithms for the optimisation of synthesis parameter values using a fitness function determined by the user.

[13] A user test conducted by the author produced results which corroborate this finding.

Within this second category, interactive genetic algorithms offer a promising means of exploring search spaces whose contours are mountainous – that is to say, where there are a number of local fitness minima and maxima. For example, the complex interaction between carrier and modulator frequencies and amplitudes means that there may be more than one good candidate solution in an FM parameter space. Mutation and crossover help to minimise the risk that the GA converges prematurely on a local maximum (or minimum).

For a parameter space which is more linear, however, and whose dimensions map more readily to acoustical attributes, it is more likely that there is (at best) only one optimum solution, and that the fitness contour of the space consists only of one peak. In such a space, the search algorithm can then be seen less as a process of *optimisation*, and more one of *localisation*.

The dimensions of the three attribute spaces which have been constructed for the work described in the next three chapters are measurable acoustical quantities whose mapping to the parameters of synthesis is linear and uncomplicated. Given this correspondence between the attribute space and parameter space, clearly the provision of a number of single dimension controllers (sliders or rotary dials), each mapped to one of the dimensions of the space would be the most straightforward. This, however, assumes, firstly, that the users are able to identify aurally individual acoustic parameters and associate them with individual controllers; and, secondly, that they are able to predict the effect of manipulating a given controller. This may be relatively easy to do in a simple attribute space, but become progressively more difficult as the dimensionality increases.

The purpose of this study, then, is to examine the operation of an alternative localisation algorithm and interface in three contrasting attribute spaces, two of which are three dimensional and one which is seven dimensional; and to compare its operation with one which simply provides controllers, each mapped to one dimension of the space.

Chapter 5 – A perceptual study of a .simple timbre space

5.1. Introduction

This chapter and the two chapters following describe the empirical work of this thesis. The search strategy which it involves is a localisation algorithm driven by iterated similarity-dissimilarity judgments made by the user. Detailed discussion of its operation, the rationale for its use and the results of user testing is deferred, however, to the next chapter; the purpose of this chapter is to describe the methodology and results of a series of listening tests designed to assess the suitability of a simple three dimensional attribute space as a testing bed for the search strategy.

The previous chapter examined a number of strategies for mapping between attribute spaces and parameter spaces. Where the mapping between an attribute space and a parameter space is complex and non-linear (as is the case when using FM synthesis, for example), optimisation algorithms which make other parts of the search space available through, for example, mutation and crossover, have been shown to be effective (Horner, Beauchamp *et al.*, 1993; Mitchell and Pipe, 2005). However, where the attribute space maps in a fairly straightforward way to the parameter space, a more direct method which converges on an optimum solution without the disruptive effects of mutation and crossover is likely to be more successful. This is the rationale for the search strategy to be discussed in the next chapter.

However, its success is entirely dependent on the ability of the user to perceive relative Euclidean distances in the space. In other words, for any three

points A, B and C disposed within an attribute space, such that the distance AC is greater than the distance BC, the difference in those distances must be reflected in perceptual judgments of timbral distance.

While the studies conducted on a number of attribute spaces (reviewed in chapter three) have demonstrated correspondences between Euclidean and perceptual distances in particular instances, it cannot, of course, be assumed that this will be generally true for all attribute spaces. In this chapter, the methodology and results of a series of listening tests, designed to test whether this is the case within a previously constructed three-dimensional attribute space, are presented.

5.2. The attribute space

The attribute space to be examined in this chapter consists of time invariant sounds i.e. they have static spectra. Using Carl Seashore's distinction between *timbre* – the spectral aspect of sound - and *sonance* - the time-variant aspects of sound (onset, vibrato, decay, spectral fluctuation etc) (Seashore, 1967), we focus here on the former.

To understand the motivation for the choice of a simple, low dimensional timbre space for the empirical work presented here, it is useful briefly to review the role of *formants* in timbre. The notion that timbre was linked to the frequency spectrum of the steady state portion of an instrumental tone was proposed in the nineteenth century, most notably by Helmholtz (1954). However, the shifting of a given spectrum up and down in frequency, preserving the amplitude and frequency ratios between its partials, nevertheless results in changes in timbre. Slawson (1968) demonstrated that a sound's *formant* characteristics provided a better model for understanding the relationship between spectrum and timbre. A

formant is a broad frequency region which causes an increase in amplitude of any spectral component partial falling within its range (Handel, 1989). Slawson, and subsequently Plomp & Steeneken (1971) demonstrated that perceived timbral similarities were more readily attributed to invariances in formant frequencies than to invariances in the overall spectral envelope. Formant terminology is more usually applied to the description of vocal systems; however, the frequency spectrum of a given instrumental sound will also have characteristic formants, which do not shift in frequency with changes in the frequency of the fundamental. Since in 'real world' acoustical systems, formant frequencies are associated with the resonance frequencies of the system (the body of a guitar or a violin, for example), Slawson, and subsequently Balzano (1986) proposed a physical model for understanding the spectral aspects of timbre, or 'sound color'.

The axes of the attribute space selected for the study reported in this chapter are formant centre frequencies; the stimuli drawn from this space sound subjectively like a collection of more or less open and closed vowel sounds. Although we are not primarily concerned with vowels as such, a simple attribute space, loosely based on vowels, has been chosen for this study; firstly, because it is simple and easily synthesizable, and secondly, because the use of such a space will allow a relatively wide range of timbral variation in the set of sounds to be generated within an otherwise very circumscribed space. The simplicity of this space will allow us to conduct the first test of the search strategy in a relatively well-understood context.

The synthesis method used to generate the stimuli for the work described both in this chapter and in chapter six and seven, is *additive synthesis* – the generation of complex waveforms by a summing of sinusoidal components according to the Fourier theorem. Based on the discussion of synthesis methods in

chapter two and in particular on the evaluation of Tolonen *et al* (1998), it has been chosen for the following reasons.

Firstly, and most importantly, the method is well-behaved - changes to the additive synthesis parameters map in a generally linear manner to timbral change. Secondly, the sounds are easily generated. One of the disadvantages of additive synthesis (again discussed in chapter two) is the amount of control data required. However, this is less of a problem for the purposes of this study, firstly because the sounds are time-invariant (which considerably reduces the amount of data required) and secondly, because the axes of the space are formant centre frequencies (metaparameters, to use Jaffe's terminology (Jaffe, 1995)), rather than the amplitudes of individual harmonics. This also addresses the problem of perceptibility identified by Tolonen *et al* (see figure 2.6 in chapter two) ; while changes in individual harmonic amplitudes do not result in significant timbral change, the perception of sound will be noticeably altered by changes in its formant structure.

5.3. Objectives

The experiment had three objectives. The first was to determine whether subjects' abilities to perceive relative distances between sounds located in the space would be significantly higher than chance performance. It was hypothesized that participants' scores would differ significantly depending on the direction in which the sounds differed (i.e. the axis along which the sounds were aligned).

The user-driven search strategy which will be proposed in chapters six and seven is dependent on candidate solutions being presented to the user which are sufficiently far apart in the space for a choice to be made between them, based on

perceived differences. Thus, the second aim was to examine the perceptual granularity of the space – specifically, the maximum distance between two sounds for which there is no difference in perceived timbre. Previous studies of the just noticeable difference or *difference limen* (DL) have explored vocal timbre spaces, with a view to establishing the accuracy necessary for the analysis and synthesis of spoken vowels (Flanagan, 1955; Flanagan, 1957; Flanagan and Saslow, 1958; Flanagan, 1972; Mermelstein, 1978; Hermansky, 1987; Gagne and Zurek, 1988; Kewley-Port and Watson, 1994; Lyzenga and Horst, 1997; Dissard and Darwin, 2001). Typically, stimuli for these studies have been synthesized vowels derived from spectrographic measurements of vowel spectra, containing three or four formants (F1, F2, F3 and F4), and which were presented to subjects as standard, together with altered versions in which the centre frequencies of a number of formants (typically F1 and F2) have been incremented/decremented. A hypothesis in our study was that, firstly, this distance, averaged out for all subjects used in the listening tests, would vary in different parts of the space; and that, secondly, such variation may account for any variations in the performances of the search strategies described in the next chapter.

Finally, the third aim was to examine the degree of correlation between the scores on the distance perception task and those on the perceptual granularity task.

5.4. Stimuli

Two sets of electronically synthesized pitched and non-pitched waveform stimuli were generated. The spectra of the pitched stimuli contained 73 harmonics of a fundamental frequency (F0) of 110 Hz, each having three prominent formants, I, II and III. The formant peaks were all of the same amplitude relative to the

unboosted part of the spectrum (20 dB) and bandwidth (Q=6) [14]. The centre frequency of the first formant, I, for a given sound stimulus, was one of a number of frequencies between 110 and 440 Hz; that of the second formant, II, was one of a number of frequencies between 550 and 2200 Hz, and that of the third, III, was one of a number of frequencies between 2200 and 6600 Hz.

The non-pitched stimuli consisted of white noise, band-pass filtered in such a way as to form spectra with formant structures as described above. Each formant had identical bandwidths (Q=10) and boost (20 dB). These stimuli were used to examine the perceptual granularity of the attribute space – noise was used rather than sounds whose spectra were discrete and harmonic (i.e. pitched) because of the confounding effect on the difference limen caused by the alignment, in some stimuli, of an individual harmonic with a formant peak (Gagne and Zurek, 1988; Kewley-Port and Watson, 1994); this problem is eliminated if the spectrum is non-discrete.

Each sound could thus be located in the three dimensional space illustrated in figure 5.1.

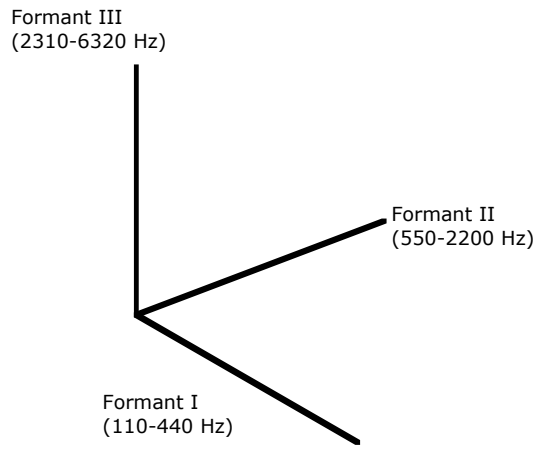


Figure 5.1: The three dimensional 'formant' space.

[14] Q is a measure of the damping of a filter or oscillator, and is given by $f_0 / \Delta f$ where f_0 is the centre frequency and Δf is the bandwidth.

All stimuli, pitched and non-pitched, were generated using Csound, and were exactly two seconds in duration, with attack and decay times of 0.4 seconds.

5.5. Test 1 - perceptual granularity

5.5.1. Procedure

We consider first the tests in which the perceptual granularity of the space was examined. Forty-eight tests were prepared, each of which consisted of a pair of noise based stimuli whose alignment in the space took one of the following forms:

- co-incident (i.e. the sounds were identical). These were used as a control.
- separated along the formant I axis by a formant centre frequency difference of Δf_1. This is expressed as a Weber ratio, equating to

  $$\frac{\Delta f_1}{f_1} * 100 = 5.95$$

  where f_1 is the lower frequency of the pair. (This figure was arrived at as a result of a pilot study, and corresponds to shifting the formant peak by about a semitone.)
- separated along the formant I axis by a formant centre frequency difference of Δf_2, expressed as a Weber ratio of

  $$\frac{\Delta f_2}{f_2} * 100 = 12.25$$

  where f_2 is the lower frequency of the pair.
- separated along the formant II axis by a Weber ratio of 5.95
- separated along the formant II axis by a Weber ratio of 12.25
- separated along the formant III axis by a Weber ratio of 5.95
- separated along the formant III axis by a Weber ratio of 12.25

None of the stimulus pairs were aligned along more than one axis; as stated above, this was in order to provide data on whether the ability to detect timbral difference was affected by the axis along which the pairs were aligned.

The attribute space was subdivided into eight subspaces **a-h**; sound pairs were located in the space so that in each subarea, there was a pair located along each axis separated by both Δf_1 and Δf_2, six pairs in all; so, for example, in area **c**, the pairs were as shown in figure 5.2.

	Lower frequency of pair	Δf_1	Δf_2
Formant I shift	310	328.43	-
Formant I shift	310	-	347.96
Formant II shift	776	822.14	-
Formant II shift	776	-	871.02
Formant III shift	2903	3075.61	-
Formant III shift	2903	-	3258.49

Figure 5.2: Disposition of tone pairs in subspace area c.

Areas **a, b, d, e, f, g and h** were correspondingly populated.

Twenty test subjects were used for this part of the study (only nineteen responses proved to be usable, however[15]). All students were in the Sir John Cass Department of Art, Media and Design of London Metropolitan University, studying either music technology or musical instrument building – consequently, these subjects were accustomed to listening critically to sound. The tests were presented through Sony MDR-V300 headphones to the test subjects in the form of a series of Web pages accessed individually from a desktop computer; an example page is shown in figure 5.3.

[15] One subject (number 5) gave two responses – 'no difference' and 'slight difference' to the same stimulus.

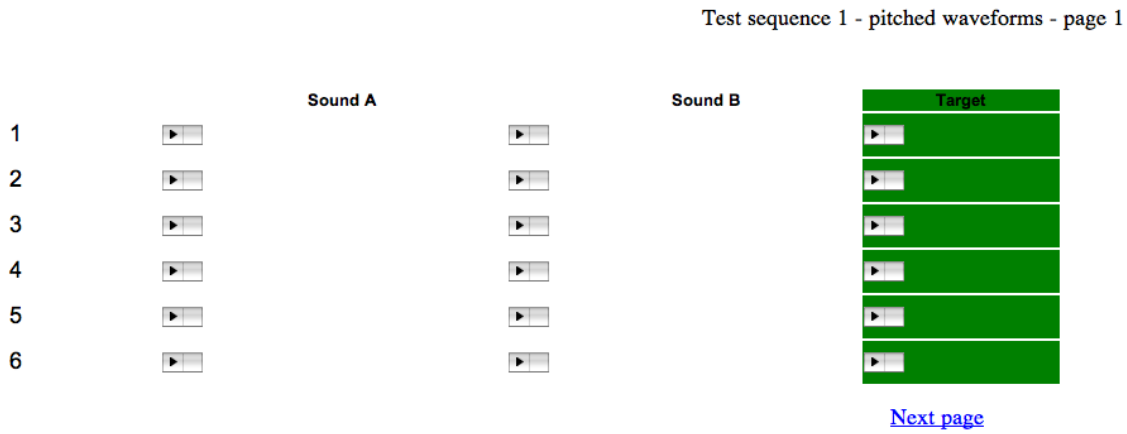


Figure 5.3: Example listening test Web page.

The subjects were divided into two groups of ten; one group received the sequence in one random order, the other group received it in another random order, in order to dilute fatigue effects. The procedure was explained, and subjects encouraged to acclimatise themselves to the sounds, and to set the headphone volume at a comfortable level. Subjects were asked to listen to each pair, and rate them for the degree of perceived difference – choices were 'no difference', 'slight difference' and 'clear difference'.

5.5.2. Results

Values of 0 were assigned to 'no difference' ratings, 1 to 'slight difference' ratings and 2 to 'clear difference' ratings. For each subject, the ratings was summed and broken down by formant - that is to say, by the axis along which the sounds in each pair were aligned, as shown in figure 5.4.

Subject	Formant I	Formant II	Formant III
1	16	28	20
2	18	26	10
3	3	22	19
4	8	32	24
6	15	30	17
7	11	27	19
8	11	30	26
9	11	28	12
10	15	29	11
11	16	27	15
12	0	15	4
13	17	26	15
14	8	22	16
15	19	25	10
16	10	29	15
17	19	27	17
18	9	27	8
19	2	15	15
20	10	22	14

Figure 5.4: Breakdown of perceptual granularity results by formant.

The data was analysed using Friedman's ANOVA (1937), which is a non-parametric repeated-measures test for ordinal data.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Formant_I	19	11.47	5.680	0	19
Formant_II	19	25.63	4.633	15	32
Formant_III	19	15.11	5.343	4	26

Ranks

	Mean Rank
Formant_I	1.37
Formant_II	2.97
Formant_III	1.66

Test Statistics[a]

N	19
Chi-Square	28.187
df	2
Asymp. Sig.	.000
Exact Sig.	.000
Point Probability	.000

Figure 5.5: Friedman's ANOVA output from SPSS.

The data shows the variance between the ratings for different axis alignments to be significant ($\chi_r^2(2, N = 19) = 28.187$, $p < .01$).

Post hoc tests were run on this data, using the Wilcoxon signed-rank test and with the significance level of .05 modified by the Bonferroni correction (0.05/3= 0.017). These showed significance in the ratings differences for formant I and II alignments ($z = -3.826$, $p < .017$), significance in the ratings differences for formant II and III alignments ($z = -3.729$, $p < .017$) and no significance in the ratings differences for formant I and III alignments ($z = -1.795$, $p = .073$).

We conclude that a shift in the centre frequency of formant II in this particular space is more salient to timbral difference perception. It is important to state at this point that this, in itself, is not surprising, and should not be regarded as evidence that the second formant of any sound is, *per se*, more salient to timbral difference perception than, say, the first or third; it can be easily attributed to the greater sensitivity of the ear in this frequency region. Similarly, the relatively low mean rating for a formant I shift can also be ascribed to the comparatively low sensitivity of the ear in the frequency region occupied by this formant.

The data was also broken down by the degree of separation between the tones in each stimulus pair (Δf_1 and Δf_2). Again, using the Wilcoxon signed-rank test, the increase in the perception of difference when the tones in the stimulus pairs were separated by a frequency difference of Δf_2 was found to be significant ($z = -3.833$, $p < .05$).

Out of the 480 tests conducted where the sound stimulus pairs were separated by Δf_1 (20 subjects x 24 tests), 28.75% of the ratings were 0 (i.e. the subjects could not hear a difference in the stimuli). As one might expect, this

figure drops to 10.625% in tests where the sound stimulus pairs were separated by Δf_2.

5.6. Test 2 - Euclidean distance perception – pitched sounds

5.6.1. Procedure

We consider now the tests designed to establish the mapping between perceptual and Euclidean distances. The Euclidean distance between two sounds **I** and **J** is defined in this study as

$$\delta(I,J) = \sqrt{\sum_n \left[\log\left(\frac{I_n}{J_n}\right)\right]^2}$$

where $\mathbf{I}_n$ and $\mathbf{J}_n$ are the coordinates of **I** and **J** on the ***n*** th axis and $n = 1 .. 3$. This particular metric is chosen because the nature of the space meets the requirements that.

$\delta(I,J) \geq 0$

$\delta(I,J) = 0$ if, and only if $\mathrm{I} = \mathrm{J}$

$\delta(I,J) = \delta(J,I)$

$\delta(I,K) \leq \delta(I,J) + \delta(J,K)$

where K is an arbitrarily placed point in the space.

Fifty six tests were compiled from the pool of pitched stimuli. Each test consisted of an equally spaced triplet of stimuli, A, B and C, forming a straight line in the attribute space, and whose alignment took one of the following trajectories:

- along the formant I axis
- along the formant II axis
- along the formant III axis

- along both the formant I and II axes
- along both the formant I and III axes
- along both the formant II and III axes
- along all three axes

Triplets were used which were aligned along only one axis, thus allowing direct comparison with the data from the granularity test reported in section 5.5. However, in order to investigate whether relative Euclidean distances could be perceived when the alignment was along more than one axis, the stimuli also included triplets which were aligned accordingly. Results from these tests were used to inform the design of a subsequent investigation (reported in section 5.8) in which each triplet had projections on more than one axis, but also did not form a straight line in the space.

The three stimuli making up each triplet were separated from each other by a frequency ratio of 1.3 – so, for example, the stimuli making up a triplet aligned along the formant II axis had identical formant I and III centre frequencies, but their formant II centre frequencies were f, 1.3f and $(1.3)^2$f. Thus, the distance from the A to C was twice that from A to B and from B to C.

The triplets were disposed in the space, such that each of the areas **a** to **h** contained seven triplets, each aligned along one of the trajectories described above.

Twenty test subjects were used for this part of the study, who were paid for their participation. Fifteen of them were music students at City University, London, the remaining five were students in the Sir John Cass Department of Art, Media and Design of London Metropolitan University, studying either music

technology or musical instrument building – consequently, these subjects were accustomed to listening critically to sound. The tests were presented through Sennheiser PX-30 headphones to the fifteen City University test subjects, and through Sony MDR-V300 to the five Sir John Cass students in the form of a series of Web pages accessed individually from a desktop computer. (The possible distorting effect of using two different models of headphones is noted, and discussed later). As before, the subjects were divided into two groups of ten, and the stimuli presented in a different random order to each group.

Each subject was asked to listen to the 56 tests, and for each of the tests, to indicate which of the first two stimuli of the triplet sounded more like the third (the standard). (The first two stimuli (i.e. A and B) of half the triplets, randomly chosen, were swapped to ensure that in approximately half the tests the first sound was actually closer to the standard.) In all cases, subjects were able to audition any sound as often as they wished, before making a decision.

5.6.2. Results

The mean subject score for all 56 tests was 37.95 (67.77%). A chi-square 'goodness of fit' test was conducted, yielding χ^2 (1, N = 56) = 6.38, $p < 0.05$ [16] (where the comparison was with chance). This is a significant result and suggests that subjects are, in general, able to perceive relative Euclidean distances between pitched sounds aligned in a straight line in this particular attribute space.

More significant, however, in the light of the data from test 1, was the variation in the number of 'correct' identifications when broken down by formant along which the ABC triplets were aligned (see figure 5.6).

[16] χ^2 is reported with Yates' correction, used when *df* (degrees of freedom) = 1

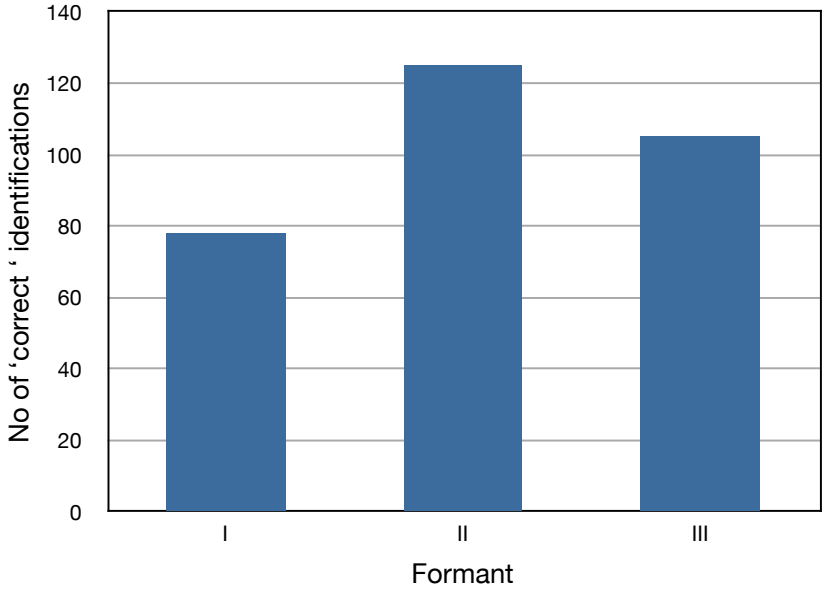


Figure 5.6: Correct identifications broken down by formant - pitched sounds.

This graph shows the number of correct identifications made by subjects broken down by formant (note that the graph excludes data from those tests where the triplets were aligned along more than one axis). A one-way repeated measures ANOVA showed the variation to be significant, $F(2, 38) = 13.26$, $p < .001$. (Maunchly's test indicated that the assumption of sphericity had not been violated, χ^2 (2)= 2.07, $p = .36$.) The null hypothesis (that the formant which is shifted (I, II or III) is not a factor in the ability of subjects to perceive relative Euclidean distances in the space) can therefore be rejected.

Post hoc tests using paired t-tests showed no significant difference between the number of correct identifications when triplets were aligned along the formant II axis and those when the alignment was along the formant III axis ($t(19) = 1.93$, $p = .069$). However, the difference was significant between formants I and II ($t(19) = 6.09$, $p < .017$) and between formants I and III ($t(19) = 2.93$, $p = .009$). [17]

[17] Again, a 'Bonferroni correction' was made to the .05 significance level (.05/3 = .017, 3 being the number of t-tests conducted).

Finally, a one-way ANOVA was conducted to establish whether the use of two different models of headphones had a confounding effect on the overall results. No significant difference was found in the mean scores from those subjects who had used the Sennheiser PX-30 and those who had used the Sony MDR-V300 ($F(1,8) = 5.30$, $p = .0502$).

To summarise: it appears that differences in Euclidean distances between three pitched sounds aligned in a straight line in this attribute space are, in general, perceptible, and the axis orientation is a significant factor in subjects' ability to perceive relative distances in the space.

5.7. Test 3 - Euclidean distance perception – non-pitched sounds

5.7.1. Procedure

In order to establish if this property of the space was confined only to pitched, harmonic spectra, or if this could be generalised to a wider set of sounds having this formant structure, a similar series of tests was conducted on nineteen subjects using a selection of noise based sounds. These sounds differed from the 'pitched' sounds only in the waveform used; the formant structure was the same in both categories of stimuli.

5.7.2. Results

The mean subject score for all 56 tests was 44.89 (80.17%). A chi-square 'goodness-of-fit' test (with Yates' correction) was conducted, yielding χ^2 (1, N=56) = 20.39, $p <$ 0.01 (where the comparison was with chance). Again, this is a significant result,

and suggests that subjects are, in general, able to perceive relative Euclidean distances between sounds aligned in a straight line in this particular attribute space. Although the number of correct identifications rose when the triplets were aligned along the formant II axis, the variance was not significant, $F(2, 36)=1.904$, $p=.164$; see figure 5.7.

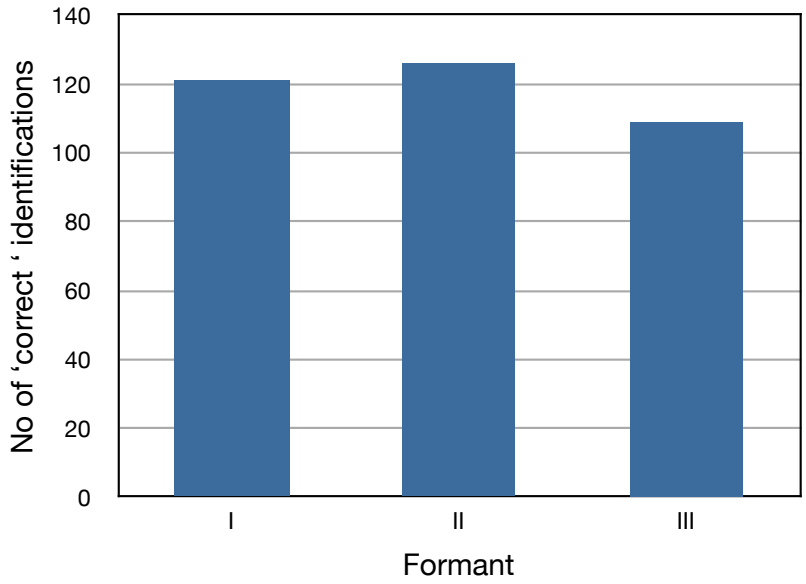


Figure 5.7: Correct identifications broken down by formant – non-pitched sounds.

We conclude that differences in Euclidean distances between three non-pitched sounds aligned in a straight line in this attribute space are also, in general, perceptible; however, there is no strong evidence to suggest, in this case, that this varies with formant alignment.

5.8. Test 4 - 'Bent line' triplets

5.8.1. Introduction

While the work described above suggests that listeners are able to perceive relative distances of sounds located on a straight line in the coordinate space, and

provides us with useful data on the extent to which the formant frequency range affects timbre discrimination, it tells us nothing about the perception of distances between sounds which are located more randomly in the space, i.e. on a bent line. In other words, whereas in tests two and three we were concerned with sounds aligned in a straight line, we now wish to consider the ability to perceive relative Euclidean distances in different directions in the timbre space. A number of further tests were therefore conducted to examine this.

5.8.2. Procedure

For each of the eight areas **a** to **h**, six tests were devised. Each test consisted of a triplet of pitched stimuli, **A**, **B** and **C,** disposed in the space such that ABC did not form a straight line, AC and BC had projections on all three axes, and the Euclidean distance AC was greater than that of BC (a ratio of AC:BC = 1.732 : 1; this is a smaller ratio than that chosen for the 'straight line' tests, but was the maximum that could accommodated within one area.) In all cases, C was the initial stimulus and A and B were the probes. The six test triplets for area **a** were constructed as follows.

Fixing the position of A at 238 Hz, 1193 Hz and 2233 Hz, three positions for C were found, such that A and C were opposite apices of a double cube whose dimensions along the formant axes I, II and III corresponded to frequency ratios respectively of (1.1911, 1.4186, 1.4186), (1.4186, 1.1911, 1.4186) and (1.4186, 1.4186, 1.1911). From these three positions for C, six positions for B were found, such that B and C were opposite apices of a cube whose dimensions along the formant axes I, II and III corresponded to a frequency ratio of 1.191 (a Weber ratio of 19.1, or 302.69 cents – just over a minor third in the tempered scale).

Thus, for one of the six tests in area **a**, sound A has centre frequency coordinates 238 Hz, 1193 Hz and 2233 Hz, sound B has centre frequency coordinates 238 Hz, 2015.76 Hz and 2659.63 Hz and sound C has centre frequency coordinates 283.47 Hz, 1692.41 Hz and 3167.77 Hz (see fig 5.8).

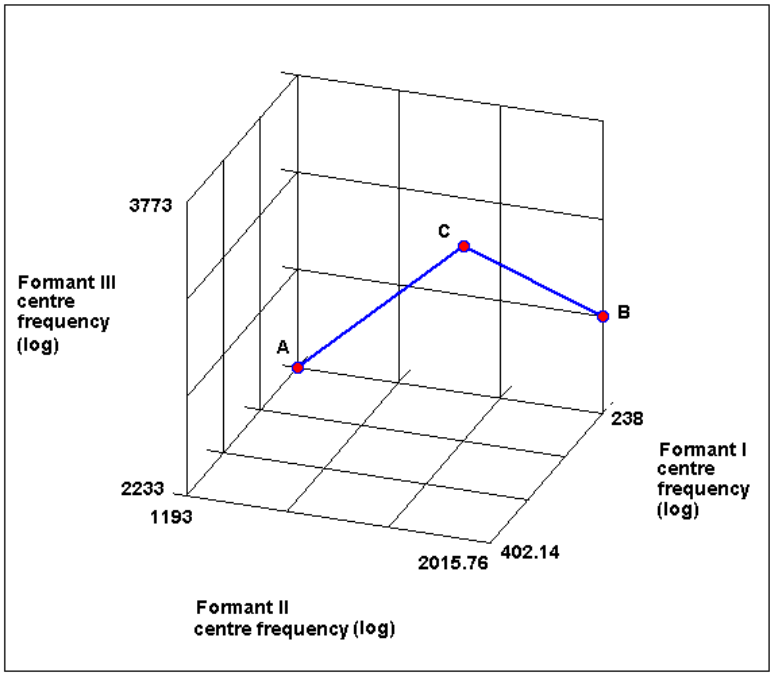


Figure 5.8: Example of 'bent line' triplet ABC.

The Euclidean distance AC is then

$$AC = \sqrt{\left[\log\left(\frac{283.47}{238}\right)\right]^2 + \left[\log\left(\frac{1692.41}{1193}\right)\right]^2 + \left[\log\left(\frac{3167.77}{2233}\right)\right]^2} = 0.227797$$

The Euclidean distance BC is then

$$BC = \sqrt{\left[\log\left(\frac{283.47}{238}\right)\right]^2 + \left[\log\left(\frac{1692.41}{2015.76}\right)\right]^2 + \left[\log\left(\frac{3167.77}{2659.63}\right)\right]^2} = 0.131518$$

which is a 1.732:1 ratio. Analogous sets of tests were generated for each of the areas **b** to **h.**

As before, twenty test subjects were used for the study; the procedure is as described in the previous tests. Fifteen subjects used Sennheiser PX-30, the remaining five Sony MDR-V300 headphones. The forty-eight tests (six tests x eight areas) were presented to half the subjects in one random order, and to the other half in another random order. Each subject was asked, for each of the tests, to indicate which of the first two stimuli of the triplet sounded more like the third. (As before, the first two stimuli of half the triplets, randomly chosen, were swapped to avoid giving clues to the subjects). In all cases, subjects were able to audition any sound as often as they wished, before making a decision.

5.8.3. Results

The mean number of 'correct' identifications for all 48 tests was 35.05 (73.02%). A chi-square 'goodness of fit' test (with Yates' correction) was conducted, yielding χ^2 (1, N=48) = 28.1, $p < 0.01$ (where the comparison was with chance).

We conclude from this that subjects are, in general, able to perceive relative Euclidean distances between three pitched sounds A, B and C in this particular attribute space, even where the vectors AB and BC are differently orientated; it is also noteworthy that the percentage of correct identifications where ABC formed a bent line exceeded the percentage of correct identifications where ABC was straight, even where the difference ratio was not as great (1.732:1 as opposed to 2:1 in test 2). There were slightly higher scores where the vectors AC and BC had a non-zero projection along the formant II axis, but not enough to be statistically significant.

Again, a one-way ANOVA was conducted to establish whether the use of two different models of headphones had had a confounding effect on the overall results. No significant difference was found in the mean scores from those subjects who had used the Sennheiser PX-30 and those who had used the Sony MDR-V300 ($F(1,8) = 1.31$, $p = .285$).

5.9. Correlation of results

Finally, the data was analysed to determine whether there was significant correlation between the scores on the distance perception task (tests 2 and 3) and those on the perceptual granularity (test 1) task.

Because each of the 48 perceptual granularity pairs were aligned along one axis only, the results could be compared only with data from those Euclidean distance tests where the triplets were similarly aligned along one axis (thus comparing like with like).

Area	Formant axis	**Perceptual granularity - test 1 (Δf_1)** Mean difference rating. no difference = 0 slight difference = 1 clear difference = 2	**Perceptual granularity - test 1 (Δf_2)** Mean difference rating. no difference = 0 slight difference = 1 clear difference = 2	**Euclidean distance perception – pitched samples - test 2** Total number of 'correct' judgments	**Euclidean distance perception – noise samples - test 3** Total number of 'correct' judgments
A	I	0.632	1.526	12	18
	II	1.105	1.842	17	16
	III	0.474	1.263	11	14
B	I	1.000	1.316	13	18
	II	1.579	1.789	15	18
	III	0.632	1.000	14	9
C	I	0.632	1.316	9	17
	II	1.211	1.737	13	17
	III	0.737	1.421	13	15
D	I	0.632	1.263	9	13
	II	1.158	1.842	14	16
	III	0.474	1.211	14	14
E	I	0.158	0.684	9	14
	II	1.474	1.842	16	14
	III	0.579	1.368	13	16
F	I	0.105	0.526	12	13
	II	1.368	1.842	18	15
	III	0.684	1.105	14	14
G	I	0.105	0.684	4	12
	II	1.421	1.947	16	16
	III	0.842	1.263	13	16
H	I	0.158	0.737	10	16
	II	1.526	1.947	16	14
	III	0.737	1.316	13	11

Figure 5.9: Comparison of perceptual granularity and relative Euclidean distance perception results.

The Spearman correlation coefficient ($\mathbf{r_s}$) can be used to establish whether scores on two variables correlate, but not necessarily linearly.

A one-tailed[18] Spearman's rank correlation test on the data from test 1 (perceptual granularity – both Δf_1 and Δf_2) against the data from tests 2 and 3

[18] A 'two-tailed' test simply seeks to determine whether there is a significant correlation between two variables; a one-tailed test determines whether the correlation is positive or negative.

(Euclidean distance perception – pitched samples and noise samples respectively) yielded the results shown in figure 5.10.

	Spearman's rank correlation coefficient (r_s)	
	Euclidean distance perception (pitched samples) – test 2	**Euclidean distance perception (noise samples) – test 3**
Perceptual granularity (Δf_1) – test 1	$r_s = 0.771$ $p < 0.001$	$r_s = 0.3855$ $p < 0.05$
Perceptual granularity (Δf_2) – test 1	$r_s = 0.6993$ $p < 0.001$	$r_s = 0.4617$ $p < 0.05$

Figure 5.10: Correlation of perceptual granularity and relative Euclidean distance perception results.

The results shows clear evidence of a positive correlation between the perceptual granularity of different parts of the space and subjects' ability to perceive relative Euclidean distances in those varying regions; put simply, subjects were more likely to make correct judgments between sounds placed in a part of the attribute space where the perceptual granularity was smallest.

5.10. Summary of results

Overall, the results can be summarised as follows.

- In 28.75% of the 480 tests conducted on stimulus pairs separated by Δf_1 (the equivalent of a semitone), subjects could hear no difference between the sounds in the pair.
- In 10.625% of the 480 tests conducted on stimulus pairs separated by Δf_2, (the equivalent of a tone) subjects could hear no difference between the sounds in the pair.

- In general, a significantly high percentage of subjects were able to correctly perceive relative Euclidean distances in this particular space; 67.77% in the case of pitched sounds aligned in a straight line in the space (test 2), 80.17% in the case of non-pitched sounds aligned in a straight line in the space (test 3), and 73.02% in the case of pitched sounds arranged in a 'bent line' in the space (test 4). It is important to note, however, that the ability of subjects to perceive relative distances in the space is errorful, and this needs to be considered when constructing a search strategy based on this.

- In tests 2 and 3, there was a higher number of 'correct' identifications where the sounds were aligned along the formant II axis, although only the variance in test 2 was statistically significant. Results from test 1 show a significantly greater ability to discriminate between two non-pitched sounds placed close together in the space where the sounds are aligned along the formant II axis.

5.11. Conclusions

We conclude, firstly, that the Euclidean distances between three sounds, A, B and C, disposed in this predefined simple attribute space, such that the distance and orientation of AC is different from that of BC, are reflected in perceptual differences. It is, of course, not being claimed that there is always a simple mapping to be made between Euclidean and perceptual distances in all attribute spaces. However, we do claim that for an attribute space where such correlation exists, the degree to which this is the case will vary with the varying perceptual granularity in different parts of the space.

For the purposes of the empirical work discussed in the next two chapters, we have demonstrated that an attribute space of this type is a suitable vehicle for testing of the search strategy. The selection of an appropriate granularity for the space is also crucial. Should the timbral difference between two adjacent sounds in the space be too great, the system will have insufficient resolution to be useful; however, too fine a resolution is likely to present performance problems.

Chapter 6 - Searching two three-dimensional spaces

6.1. Introduction

The aim of the study presented in this chapter is to report on the rationale, design and operation of a strategy for searching an attribute space. The strategy makes use of an adapted *weighted centroid localisation* (WCL) algorithm, in which user input to a user/system dialog iteratively updates a probability network, which in turn steers the convergence of the candidate solution onto a 'best fit' solution. Detailed discussion of the strategy, together with contextualisation and the rationale for its use, is presented in section 6.3.2. The results of this work are presented in section 6.6 and discussed in section 6.7.

The WCL search strategy was tested in the form of two computer programs. The first of these, henceforth referred to as WCL-2, employs a forced choice similarity test, where the user is iteratively asked to judge which of two probe sounds, taken from the attribute space, more closely resembles a target sound, also taken from the space. A candidate sound is generated on each iteration.

The second variant, which we will call WCL-7, offers seven probe sounds at each stage, but otherwise operates in the same way – the subject is asked to judge which of the seven probes more closely resembles the target sound. In both cases, the candidate solution is not directly specified by the subject, but instead generated by the software based on user input.

In order to have a baseline against which to assess the success of the WCL strategy, another program was developed which provided the user with the means of manually navigating the attribute space. The user interface afforded direct access to the attribute space via individual sliders which control navigation along each axis. This is a form of *multidimensional line search* (MLS). In this strategy, the candidate sound is generated by the subject by direct navigation using the axis sliders. Further discussion, together with a rationale for this approach, is given in section 6.3.1.

To summarise: in all three strategies - WCL-2, WCL-7 and MLS - subjects were presented with a target sound taken from the attribute space, and were asked to drive a candidate sound through the space, such that it converged with the target. In all cases, the candidate solution is modified, either by the user (in the case of MLS) or by the software (in the case of WCL-2 and WCL-7) as the interaction proceeds, and it is the trajectory of the candidate solution through the attribute space which is of interest to this study in all three cases. A successful interaction is defined here as one in which there is an overall convergence of the candidate solution on the target - the most successful being characterised by the steepest gradient.

The use of a target-oriented methodology raises a number of issues, however. The first of them – that the process of timbral design and editing is not necessarily and exclusively target-oriented, but can involve exploratory and improvisational modes of usage as well - has already been considered in chapter two (section 2.6.3.2.1), but applies with equal force here. Secondly, in all the work presented in this chapter and the next, the target sound is a given, whereas when a subject is employing these, or any other synthesis algorithms in a target-oriented 'real world' situation, the target sound is likely to be imaginary, existing only in

the subject's head. The extent to which we can draw generalised conclusions from these results on the usefulness of any of these three strategies in such a real world situation is dependent on two assumptions. Firstly, it is assumed that an imagined sound is one that actually exists in the space (i.e. one that can be reached). Secondly, it is assumed that an imagined sound remains stable; that is to say, the sound which the subject is trying to create using a synthesis method remains the same. Consideration of the efficacy of these, or any other search strategies when the goal is a moving target – that is to say, the user changes his/her mind about the sound to be created - is outside the scope of this study.

In order to analyse and compare the operation of these strategies, a series of user tests was conducted. Two attribute spaces were constructed, in order to compare the operation of the three strategies in different environments, and to assess the extent to which the findings could be generalised. These were as follows:

- a three dimensional attribute space, which we will call the *formant space*, consisting of time-invariant pitched sounds, and based on the space tested and discussed in the previous chapter ;

- a three dimensional attribute space, which we will call the *SCG-EHA space*, consisting of time-variant sounds and derived from the work of Caclin, McAdams, Smith and Winsberg (2005) (discussed in chapter three).

Before discussing the WCL search strategies, the two spaces in which they operate will now be considered in detail.

6.2. Attribute spaces

6.2.1. Formant space

6.2.1.1. Rationale for the use of this space

The use of formant frequencies as the axes of the space is based on the work of Slawson (1985); as stated above, this attribute space is derived entirely from that examined in the previous chapter, which demonstrated that relative Euclidean distances in the space are reflected in perceptual judgments; more specifically, that where three sounds A, B and C are disposed in the space such that the distance AC is greater than the distance BC, subjects will perceive C as being timbrally more similar to B than to A. This makes the space a suitable vehicle for testing a search strategy driven by similarity/dissimilarity judgments.

6.2.1.2. Construction of the attribute space

The characteristics of the space and of the sounds inhabiting it are the same as those of the space discussed in chapter five (sections 5.2. and 5.4); they are recapitulated here for clarity.

The sounds inhabiting this space were all exactly two seconds in duration, with attack and decay times of 0.4 seconds. Their spectra contained 73 harmonics of a fundamental frequency (F0) of 110 Hz, each having three prominent formants, I, II and III. The formant peaks were all of the same amplitude relative to the unboosted part of the spectrum (20 dB) and bandwidth (Q=6). The centre frequency of the first formant, I, for a given sound stimulus, was one of a number of frequencies between 110 and 440 Hz; that of the second formant, II, was one of a number of frequencies between 550 and 2200 Hz, and that of the third, III, was one

of a number of frequencies between 2200 and 6600 Hz. Each sound could thus be located in the three dimensional space illustrated below.

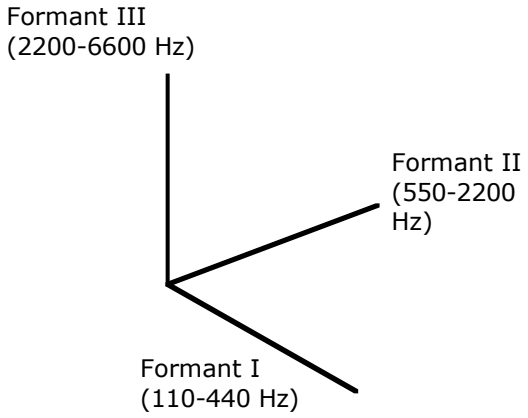


Figure 6.1: Formant attribute space.

6.2.2. SCG-EHA space

6.2.2.1. Background

The second of the two spaces is derived from the work of Caclin *et al* (2005); because of the importance of their study to the design of this particular attribute space, the motivation, methods and findings of the paper are summarised here.

The paper takes as a baseline the work of a number of researchers who have concluded from MDS studies that spectral centre of gravity (SCG) and attack time are salient acoustical correlates of timbre perception (Grey, 1977; McAdams, Winsberg *et al.*, 1995; Marozeau, de Cheveigne *et al.*, 2003). Other correlates that have been identified include spectral flux (variance of the spectrum over time) (McAdams, Winsberg *et al.*, 1995), and spectral irregularity (Krimphoff, McAdams *et al.*, 1994). Noting that, given the wide range of possible acoustical correlates, 'one can never be sure that the selected parameters do not merely covary with the true underlying parameters', Caclin *et al* performed a set of experiments to

confirm these earlier findings, by constructing spaces of synthetic sounds that varied *only* by these parameters, and conducting dissimilarity rating listening tests on those sounds. The hypothesis was that, if these correlates were correct, there should be a good match between the physical (attribute) space and the perceptual space.

Of the three different spaces generated in the study, the third (the parameters of which are described below) provided a good match. The sixteen sounds used were synthetically generated pitched tones with a fundamental of 311 Hz (E4), containing 20 harmonics.

The first variable parameter was **attack time**; the attack envelope was linear, and varied logarithmically between 15 and 20 milliseconds (it has been noted that the logarithm of attack time appears to explain the corresponding timbre dimension better than the attack time itself (Krimphoff, McAdams *et al.*, 1994; McAdams, Winsberg *et al.*, 1995))

The second was **spectral centre of gravity** (SCG), or spectral centroid, defined here as the amplitude-weighted mean frequency of the energy spectrum. It has been noted that this parameter corresponds to the perception of brightness in the sound. For all stimuli, the amplitude A_n of any harmonic n was calculated by

$$A_n = k \times 1/n^{\alpha}$$

where *k* is an arbitrary value and α a value determined by the SCG. This, in turn, is given by

$$SCG = \frac{\sum_n n \times A_n}{\sum_n A_n}$$

and is the harmonic rank number where the SCG is located. Examples of spectra with high and low SCGs are illustrated in figure 6.2 (amplitudes are given on arbitrary linear scales). The SCG varied in linear steps between 3 to 4.5 in harmonic rank units - that is to say, between 933 and 1400 Hz.

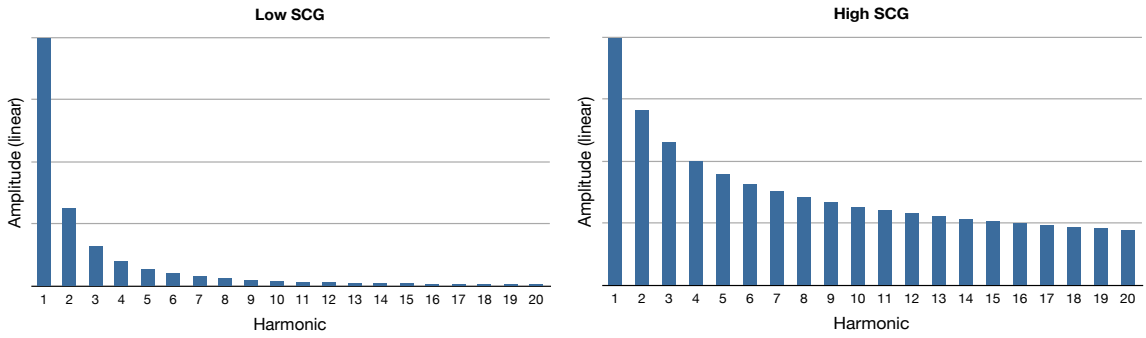


Figure 6.2: Spectra with a) low and b) high spectral centroids.

The third and last parameter was the attenuation of even harmonics relative to odd harmonics (EHA), as illustrated in the example spectra in figure 6.3. The attenuation of even harmonics ranged from 0 (as in the first example spectrum) to 8 dB, and could take 16 different values, separated by equal steps.

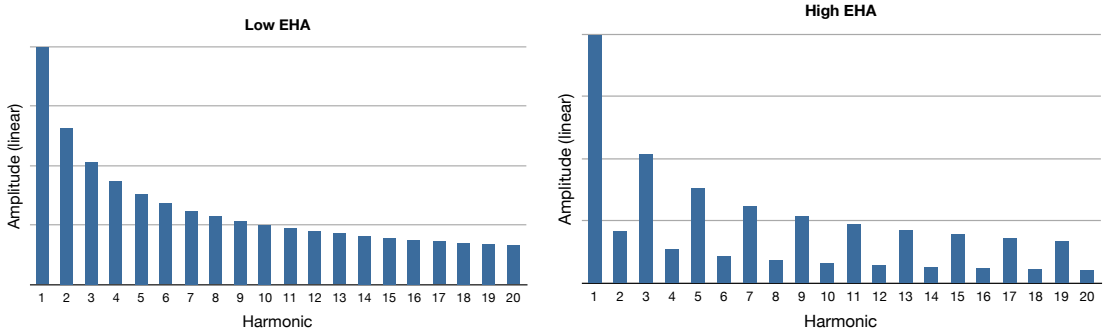


Figure 6.3: Spectra with a) low and b) high even harmonic attenuation.

In all cases, the amplitude envelope always consisted of a linear rise time (varied as described above), followed by a plateau and an exponential decay of 200 milliseconds. For this three dimensional space, the authors found a good match between the perceptual and physical dimensions. Of particular interest was

the finding that there were two latent classes of subjects – subjects in one class weighted the dimension corresponding to EHA more heavily than the other two parameters, whereas subjects in the other class weighted more heavily the dimension corresponding to SCG; in other words, in broad terms, some listeners discriminated between sounds based on EHA, others on the basis of SCG.

6.2.2.2. Rationale for the use of this space

This space is a suitable vehicle for testing for our purposes because

- a good mapping between physical and perceptual dimensions has been found in this particular space - this means that no listening tests of the type described in chapter five need be conducted; and

- sounds in the space vary only by their attack time, SCG and EHA - this means that they are easily synthesizable and the search will not be disrupted by timbral variations (specificities) which are not accounted for by the three axes.

6.2.2.3. Construction of the attribute space

An attribute space whose dimensions were those of the study described above was constructed. The sounds inhabiting this space were pitched with a fundamental of 311 Hz, and contained 20 harmonics. The attributes dimensions were as follows:

- Rise time, ranging from 0.01 to 0.2 seconds in 11 logarithmic steps. In all cases, the attack envelope was linear.
- EHA - attenuation of even harmonics relative to the odd ones in the range 0 dB to 10 dB – again in 11 steps.

- SCG - spectral centroid, in the range 3.000 to 8.000, in 15 linear steps. This corresponds to a spectral centroid range of 933 Hz to 2488 Hz.

Each sound could thus be located in the three dimensional space illustrated in figure 6.4.

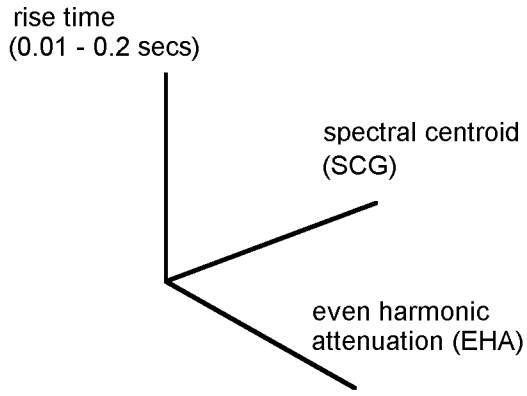


Figure 6.4: SCG-EHA attribute space: axes are rise time, degree of even harmonic attenuation and spectral centre of gravity (spectral centroid).

All sounds were generated using Csound, and normalised to –3 dB relative to full amplitude using an audio editor.

While the space is broadly similar to that in Caclin *et al*, there are two small, but important differences, which we will now consider. Firstly, the range covered on the SCG axis is wider than that in the Caclin *et al* study. This was in order to provide a greater degree of timbral variation than was apparent in that space. The range of EHA was also expanded for the same reason. The number of steps on these axes was also chosen to ensure detectable timbral difference between

adjacent discrete steps on the axes. We are free to do this because the purpose of this study is not to examine the psychoacoustical properties of the space itself, but to compare the operations of three search strategies on a given space, where the mapping between physical and perceptual dimensions has already been broadly demonstrated.

Secondly: it was noted, when constructing the space, that a change in EHA, brought about by attenuation of even harmonics, also resulted in a small change in SCG. It could be argued that the axes of the space are not, for this reason, entirely orthogonal. It is not clear in Caclin *et al* how the authors dealt with this. In the present study, the amplitude reduction of even harmonics is accompanied by a compensating **increase** in the amplitudes of odd harmonics, thus preserving SCG while maintaining variation in EHA.

We do not believe that these small modifications to the space make significant alterations to its psychoacoustical properties.

6.3. Search strategies

We turn now to the discussion of the WCL-2, WCL-7 and MLS search strategies, beginning with MLS.

6.3.1. Multidimensional line search (MLS)

As noted above, this method provides the subject with sliders giving direct access to the axes of the space, as shown in figure 6.5.

Figure 6.5: Multidimensional line search in a three dimensional space using three sliders.

Thus, adjusting the position of the first slider, for example, would have the effect of changing the first formant centre frequency in the case of the formant space, and of changing the SCG of the sound in the SCH-EHA space.

This strategy has the virtue of simplicity; indeed, for a space of low dimensionality, it may be the most effective, fulfilling two of the most important criteria proposed by Jaffe (parameter changes have a perceptible effect and are 'well-behaved' rather than 'wildly non-linear') (Jaffe, 1995). However, a successful interaction using this method is entirely dependent on the ability of the user to a) hear the individual parameters being modified and, crucially, to understand the aural effect of changing any one of them. [19]

6.3.2. Weighted centroid localisation (WCL)

6.3.2.1. Introduction

Chapter two (section 2.3) defined the user interface, both generally and in the context of sound synthesizers. However, the WCL interface presented here is

[19] It is, however, the same assumption made in the interface design of conventional subtractive synthesizers.

different. In contrast to the MLS interface (and the user interface characteristic of the standard hardware and software synthesizer), the WCL interface offers no means of directly accessing the synthesis parameters. Both the rationale for, and the implications of this will be discussed now.

A number of differing types of user interface architectures were explored and contrasted in chapter two. In order to provide the means by which the user formulates and completes an editing task, all of these architectures represent the sound in some visual way - as a waveform or frequency spectrum, in terms of its synthesis parameters, expressed numerically, or as a network of functional components. Section 2.6.3.2.1 noted that the typical synthesizer interface can be better characterised as indirect manipulation (see figure 2.13), in that there are two levels of feedback; firstly, from the interface representation (the appearance of the waveform at any given moment, for example, or the connections between components), and secondly, from the sound itself (how it sounds).

In the interface presented here, by contrast, the intervening feedback level is removed; the user is engaging with the sound as heard, rather than with a visual representation of it. Attention is focussed on the sound, and the 'interface' element is correspondingly diminished and made transparent; controls are confined to providing the means of making choices between different candidates. This has advantages, precisely because the user is engaging directly with the sound, and does not need to have an understanding of the acoustical attributes of the sound or the synthesis engine that is generating it. The subject of visual representation will be revisited, however, in the final chapter of this thesis (section 8.4.2).

Similarity-dissimilarity forced-choice listening tests typically present subjects with a pair of stimuli, and ask them to rate their degree of similarity on a numerical scale of values. It is argued here that, where such tests demonstrate a correspondence between Euclidean and perceptual distances in a given attribute space, a similar and complementary process can be used as a user-driven method for the localisation of a sound chosen from that space. In essence, the subject is presented at each stage of the interaction with a number of probes (two and seven for the WCL-2 and WCL-7 versions respectively) and asked to make a judgment as to which one of the probes most resembles an unchanging target sound. The subject's response updates a probability table, which, in turn, is used to generate a candidate solution. Over the course of the interaction, the candidate solution converges on the target.

The reason for the use of two versions of the strategy (one using two probes, the other using seven) was to observe the effect on the interaction (if any) of varying the number of choices offered. Clearly, two probes is the minimum number of choices that can be offered. The other end of the scale was taken to be seven; the assumption was made in the work presented here that meaningful comparison by subjects of the timbral qualities of more than seven probes would be prohibitively difficult. Even with seven probes, there is a significant increase in the cognitive load on the subject. It is clear that making comparisons of two probes and a target is more straightforward and easier to accomplish than comparisons of seven probes and a target. The question which was addressed in this experiment was whether this would be a significant factor in the rate of convergence with the target.

Before considering the methodology used here in greater detail, and because the strategy makes use of a probability table and a set of probes driving a weighted centroid localisation process, it is instructive to compare and contrast this process with, firstly, Bayesian methods (which make use of probability networks) and secondly, WCL methods used in other domains.

6.3.2.2. Related search methods

6.3.2.2.1. Bayesian networks

An observation made in the empirical work described in chapter five is that perception of relative Euclidean distances in a given attribute space is errorful; the "correct" judgment of relative Euclidean distances in a given attribute space is, on average, 70%. It follows that a successful search strategy driven by such judgments necessarily needs to handle a degree of uncertainty.

The use of Bayesian, or belief, networks for the representation and solving of decision problems where there is uncertainty is well established. A Bayesian network is a means of representing a set of connecting probabilities in some application domain (Pearl, 1988). Nodes in the network represent variables relevant to the particular domain; the arcs or connections between the nodes represent probabilistic relationships between those variables. Updating a node in the network with evidence results in updates being propagated to other connected nodes across the network. Bayesian networks have been successfully used in diagnosis (Andreassen, Woldbye *et al.*, 1987; Breese, Horvitz *et al.*, 1992; Heckerman, Horvitz *et al.*, 1992), forecasting (Abramson, 1994; Gu, Peiris *et al.*, 1994) , machine vision (Levitt, Agosta *et al.*, 1990), and manufacturing (Nadi,

Agogino *et al.*, 1991) (all cited in (Heckerman, Mamdani *et al.*, 1995) . More recently, Bayesian network methods have been used in audio-visual speech recognition (Choudhury, Rehg *et al.*, 2002; Nefian, Liang *et al.*, 2002) and machine listening and transcription of recorded music (Kashino and Murase, 1998; Kashino, Nakadai *et al.*, 1998; Raphael, 2002).

Bayesian methods are essentially classifiers – that is to say, they provide a mapping from some feature space, which may be continuous or discrete, to a set of discrete labels. Classification, by its nature, requires that the number of labels is significantly less than the number of objects to be classified. The WCL strategy however, does not perform a classification task in this sense; the number of possible outcomes to the search is, at the outset, equal to the number of sounds in the space. Consequently, strict Bayesian methods are not well suited to this particular problem – at best, the strategy can be described as quasi-Bayesian.

6.3.2.3. Weighted centroid localisation

Weighted centroid localisation is used in wireless sensor networks for pinpointing the position of individual sensors within the network. Before briefly describing an application of the technique, a definition and explanation is given here.

The centroid of a surface or body is its centre of mass, assuming uniform density; the centroid of a set of points in n-dimensional space is the arithmetic mean of all the coordinates of the space. Thus, the centroid of seven points in a two dimensional space whose i and j coordinates are as shown in figure 6.6.

	i	j
	4.5	0.5
	2.9	3
	5	4
	5.5	3
	2.1	1
	1	2
	7	1
	i_c	j_c
Mean	**4**	**2.1**

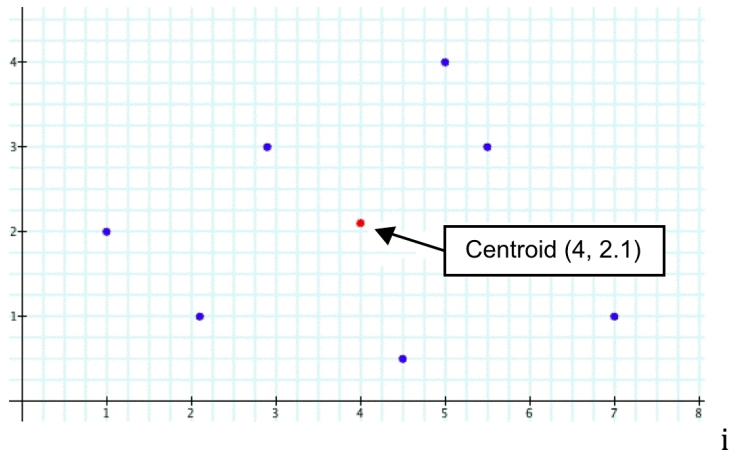


Figure 6.6: Centroid of a set of points.

The centroid of a 3 x 3 matrix is the element whose coordinates are (2,2) - that is to say, the centre of the matrix. However, if we include in the calculation of the centroid the values (or weights) of the matrix elements, we derive the weighted centroid .

The *i* coordinate of the weighted centroid is derived by

- multiplying the value (weight) of each element by its *i* coordinate, and summing the results
- dividing this by the total of all the values, or weights, in the matrix.

Similarly, the *j* coordinate of the weighted centroid is derived by

- multiplying the value (weight) of each element by its *j* coordinate, and summing the results
- dividing this by the total of all the values, or weights, in the matrix. (See figure 6.7.)

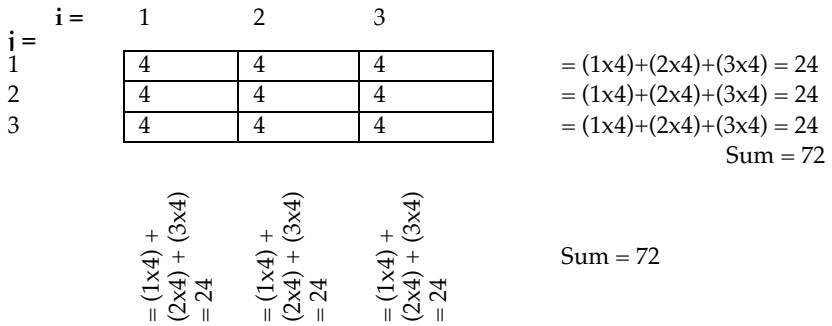


Figure 6.7: Weighted centroid of a simple unweighted 3 x 3 matrix.

The total weight of the matrix is 36 – thus, the *i* coordinate of the weighted centroid is 72/36= 2, and the *j* coordinate of the weighted centroid is also 72/36= 2.

If, however, the weight of one of the elements of the matrix is increased, the total weight of the matrix becomes 40; both the *i* and *j* coordinates of the weighted centroid becomes 2.1, shifting its location toward the element with the highest weight, as shown in figure 6.8.

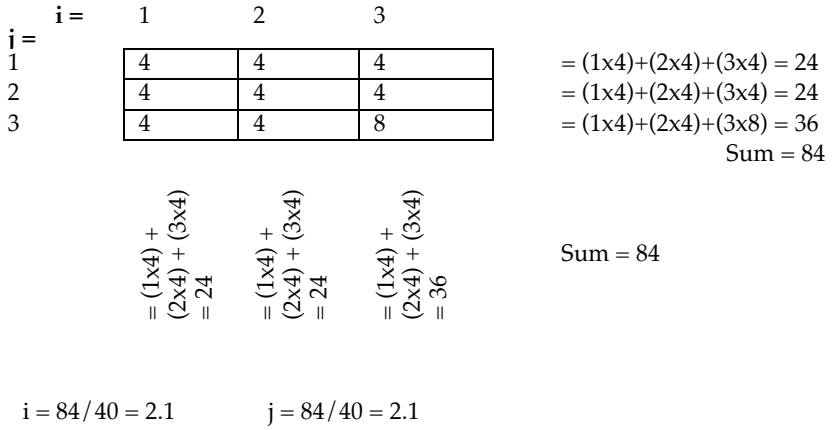


Figure 6.8: Weighted centroid of a simple weighted 3 x 3 matrix.

The notion of a weighted centroid has a number of applications and has been shown to be effective in locating individual sensors within wireless sensor networks (Blumenthal, Grossmann *et al.*, 2007). Such networks consist of a number

of nodes, some of which can determine their own positions (beacons) and others (sensors) which cannot, and which calculate their own positions by a centroid determination from the positions of the beacons in range. The weighting depends on the distance and on the characteristics of the sensor node's receivers.

The search strategy described in the following section works in an analogous way. The weighting, however, works differently; it arises from a table of cells, each corresponding to one candidate timbre in the attribute space, and whose value reflects the probability that the corresponding timbre is the target. We now turn to discussing the strategy in greater detail.

6.3.2.4. The WCL search strategy method

The WCL method is an iterated user/system dialog designed to steer a system-generated candidate sound **C** towards a goal, or target sound **T,** within the two attribute spaces described earlier, with the aim of minimising the Euclidean distance **CT**.

In general, an **n**-dimensional attribute space is constructed (such as those described earlier in this chapter), which contains, at any time, a fixed target sound **T** and a number of iteratively generated probe sounds. In addition, we construct an **n**-dimensional table, such that for each element **s** in the attribute space, there is a corresponding element **p** in the probability table. The value of any element **p** represents the probability, at any given moment, that the corresponding element **s** is the target sound- i.e. **P(s=T)**, based on information from the user.

On each step of the user/system dialog, the user is presented with the target sound T and a number of probes, and asked to judge which of the probes

most closely resembles T. This information is used by the system to generate a new candidate sound C, whose coordinates are, at any time, those of the **weighted centroid** of the probability table.

In sections 6.3.2.5. and 6.3.2.6., we consider the operation of two versions of this strategy as applied to a discrete three dimensional attribute space **S** (such as those described above), each of whose cells $\mathbf{s}_{i,j,k}$ represents a sound, and each of whose axes **1. I**, **1. J** and **1. K** represent a variable acoustical attribute of that sound. The sounds vary *only* by these attributes. The space therefore contains **I*J*K=N** cells.

A corresponding three dimensional probability table **P** was constructed. Each cell $\mathbf{p}_{i,j,k}$ in **P** holds a value reflecting the probability that the corresponding sound in **S** is the target sound – thus, there is a 1 to 1 mapping between any cell $\mathbf{p}_{i,j,k}$ and $\mathbf{s}_{i,j,k}$. At the outset, the values of all cells in **P** are initialised to 100.

6.3.2.5. WCL-2 - two-alternative forced choice

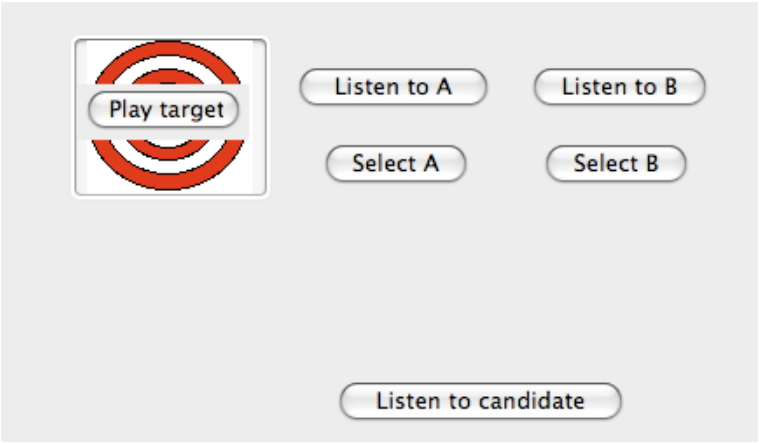


Figure 6.9: WCL-2 – interface for the two-choice algorithm.

Three sounds, chosen randomly from the space, are presented to the subject - a target sound **T** and two probes **A** and **B;** their coordinates in the attribute space are:

$A = s_{i_A, j_A, k_A}$

$B = s_{i_B, j_B, k_B}$

$T = s_{i_T, j_T, k_T}$

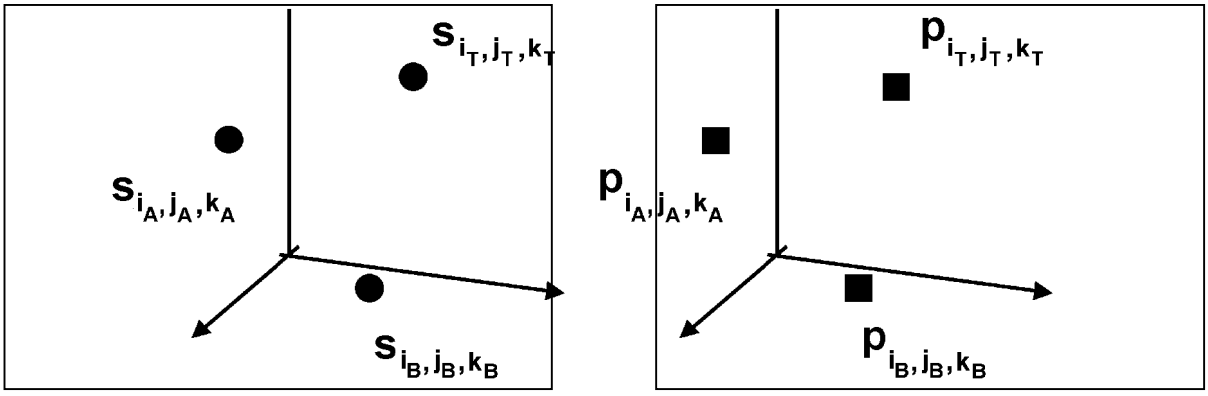


Figure 6.10 (a): Attribute space **S** (b): Probability table **P.**

Figures 6.10(a) and (b) show the three sounds **A**, **B** and **T** in the attribute space **S**, and the corresponding cells in the probability table **P**.

On each iteration of the algorithm, the subject is asked to judge which of the two probes **A** or **B** more closely resembles **T**. The subject having made a choice, the following steps are executed:

1. If **A** has been chosen, the values of all cells in **P** whose Euclidean distance from **B** is greater than their distance from **A** are multiplied by a factor of √2; the values of all other cells are multiplied by a factor of 1/√2. Justification of the choice of value for the multiplying factor is deferred to section 6.3.2.5.1.

2. If **B** has been chosen, the values of all cells in **P** whose Euclidean distance from **A** is greater than their distance from **B** are multiplied by a factor of √2; the values of all other cells are multiplied by a factor of 1/√2.

Thus, on each iteration, **P** is effectively bisected, as shown in figure 6.11, by a line which is perpendicular to the line **AB**.

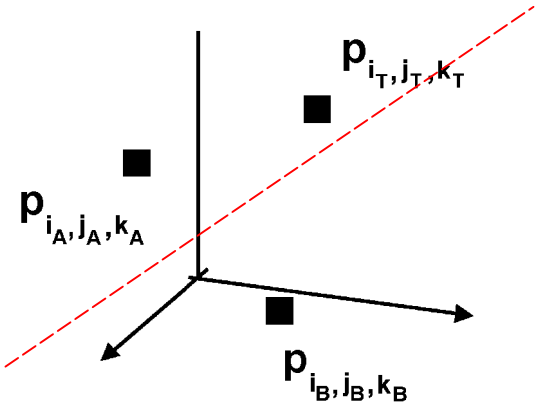


Figure 6.11: Bisection of probability table **P.**

3. Recalculate the weighted centroid **C**, as follows:

$$i_C = \frac{\sum_{x=1}^{N} w_x i_x}{\sum_{x=1}^{N} w_x}, j_C = \frac{\sum_{x=1}^{N} w_x j_x}{\sum_{x=1}^{N} w_x}, k_C = \frac{\sum_{x=1}^{N} w_x k_x}{\sum_{x=1}^{N} w_x}$$

where $\mathbf{i}_C$, $\mathbf{j}_C$, $\mathbf{k}_C$ are the coordinates of the weighted centroid **C**, $\mathbf{i}_x$, $\mathbf{j}_x$ and $\mathbf{k}_x$ are the coordinates of the **x** th cell in **P**, $\mathbf{w}_x$ is the value of the **x** th cell in **P**, and **N** is the total number of cells in **P**.

The probability space **P** having been updated, two new probes $\mathbf{A}_{new}$ and $\mathbf{B}_{new}$ are generated, and the process repeated. The generation of coordinates for $\mathbf{A}_{new}$ and $\mathbf{B}_{new}$, although pseudo-random, is nevertheless performed under the following constraints.

- $\mathbf{A}_{new}$ and $\mathbf{B}_{new}$ should be sufficiently far apart in the attribute space for there to be significant difference in their timbres. The distance was determined from data acquired from the listening tests described in chapter five. Earlier

pilot tests failed because subjects could not, in some cases, hear any real difference between the two, and therefore could not make a judgement about their degrees of respective similarity to **T**.

- A line connecting $\mathbf{A}_{\mathbf{new}}$ and $\mathbf{B}_{\mathbf{new}}$ should be more or less orthogonal to a line connecting **A** and **B**. This is to ensure that the information accumulating in the probability table **P** builds up along more than one dimension.

The algorithm in its entirety can be stated more formally as shown in figure 6.12. The function *d(x,y)* is a function which returns the Euclidean distance between *x* and *y* in the probability space **P**.

1. Generate probes *A* and *B*
2. Input *response*
3. **While** *response* $\neq$ quit **do**
   3.1. $selected\_probe \leftarrow response$
   3.2. **If** $selected\_probe = A$ **then**
      3.2.1. **For** $i = 1$ **to** I **do**
         3.2.1.1. **For** $j = 1$ **to** J **do**
            3.2.1.1.1. **For** $k = 1$ **to** K **do**
               3.2.1.1.1.1. **If** $d\left(p_{i,j,k}, p_{i_B,j_B,k_B}\right) > d\left(p_{i,j,k}, p_{i_A,j_A,k_A}\right)$ **then**
                  3.2.1.1.1.1.1. $p_{i,j,k} \leftarrow p_{i,j,k} * \sqrt{2}$
                  else
                  3.2.1.1.1.1.2. $p_{i,j,k} \leftarrow p_{i,j,k} * 1/\sqrt{2}$

      else
      3.2.2. **For** $i = 1$ **to** I **do**
         3.2.2.1. **For** $j = 1$ **to** J **do**
            3.2.2.1.1. **For** $k = 1$ **to** K **do**
               3.2.2.1.1.1. **If** $d\left(p_{i,j,k}, p_{i_A,j_A,k_A}\right) > d\left(p_{i,j,k}, p_{i_B,j_B,k_B}\right)$ **then**
                  3.2.2.1.1.1.1. $p_{i,j,k} \leftarrow p_{i,j,k} * \sqrt{2}$
                  else
                  3.2.2.1.1.1.2. $p_{i,j,k} \leftarrow p_{i,j,k} * 1/\sqrt{2}$

   3.3. $weight\_total \leftarrow 0$
   3.4. $weighted\_coordinate\_total_i \leftarrow 0$
   3.5. $weighted\_coordinate\_total_j \leftarrow 0$
   3.6. $weighted\_coordinate\_total_k \leftarrow 0$

   3.7. **For** $i = 1$ **to** I **do**
      3.7.1. **For** $j = 1$ **to** J **do**
         3.7.1.1. **For** $k = 1$ **to** K **do**
            3.7.1.1.1. $weight\_total \leftarrow weight\_total + p_{i,j,k}$
            3.7.1.1.2. $weighted\_coordinate\_total_i \leftarrow weighted\_coordinate\_total_i + \left(p_{i,j,k} * i\right)$
            3.7.1.1.3. $weighted\_coordinate\_total_j \leftarrow weighted\_coordinate\_total_j + \left(p_{i,j,k} * j\right)$
            3.7.1.1.4. $weighted\_coordinate\_total_k \leftarrow weighted\_coordinate\_total_k + \left(p_{i,j,k} * k\right)$

3.8 $i_C = weighted\_coordinate\_total_i / weight\_total$
3.9 $j_C = weighted\_coordinate\_total_j / weight\_total$
3.10 $k_C = weighted\_coordinate\_total_k / weight\_total$

3.11 Generate probes *A* and *B*
3.12 Input *response*

Figure 6.12: WCL-2 – two choice algorithm – pseudocode.

As **P** is progressively updated, its weighted centroid **C** starts to shift (at the outset, because all the cells in **P** have the same value, the centroid is located exactly at the centre of the space). If all, or most, of the subject responses are correct (i.e. the subject correctly identifies which of **A** or **B** is closer to **T**), the position of **C** progressively approaches that of **T**.

As already stated, the search strategy is user driven; thus, the subject determines when the goal has been achieved. At any point, the subject is able, by clicking on the 'Listen to candidate' button, to audition the sound in the attribute space corresponding to the weighted centroid **C**; the interaction ends when the subject judges **C** and **T** to be indistinguishable.

Data pertaining to the iteration is logged by the software – in particular, the successive positions of **A**, **B** and **C**, the degree to which the interaction has been successful is measured by the gradient of the approach of **C** to **T**, and the number of iterations required.

6.3.2.5.1. Rationale for multiplication factor value

Central to the weighted centroid localisation strategy is the process by which the probability table corresponding to the attribute space is updated. In the WCL-2 strategy, the values contained in the probability table cells corresponding to those sounds which are closer to the chosen probe in the attribute space are increased by a factor of √2 (as stated in section 6.3.2.5); the values of all other cells are multiplied by $1/\sqrt{2}$ (i.e. decreased). The rationale for this value was that $1/\sqrt{2}$

= 0.707, which corresponds roughly to the proportion of correct judgements (73.02%) made in the listening tests summarised in section 5.10 of chapter five. This will be revisited and discussed further in chapter eight.

6.3.2.6. WCL-7 - seven-alternative forced choice

We turn now to consider the second version of the WCL strategy.

Figure 6.13: WCL-7 – interface for the seven choice algorithm.

A target sound **T** and seven probes **A…G** , chosen randomly from the space, are presented to the subject; their coordinates in the attribute space are

$$A = s_{i_A, j_A, k_A}$$
$$B = s_{i_B, j_B, k_B}$$
$$C = \ldots \text{ etc}$$
$$T = s_{i_T, j_T, k_T}$$

On each iteration of the algorithm, the subject is asked to judge which of the seven probes **A** to **G** more closely resembles **T**. The subject having made a choice, the following steps are executed:

1. For each cell in the probability table **P**, establish its Euclidean distance **d** from the cell corresponding to the selected probe, and multiply its value by 100/**d**. In effect, the value of a cell increases in inverse proportion to its distance from the selected probe. Justification for this choice of factor value is deferred to section 6.3.2.6.1.

2. Recalculate the weighted centroid **C**, as described in 4.2.3.1.

3. Generate a new set of probes **A .. G**. As before, this is not entirely random, as the Euclidean distance between the probes needs to be of a sufficient magnitude to allow audible timbral differences to be perceived by the subject.

The algorithm in its entirety can be stated more formally as follows:

1. Generate probes *A* ... *G*
2. Input *response*
3. **While** *response* ≠ quit **do**
   3.1. $selected\_probe \leftarrow response$
   3.2. **For** $i = 1$ **to** I **do**
    3.2.1. **For** $j = 1$ **to** J **do**
        3.2.1.1. **For** $k = 1$ **to** K **do**
      3.2.1.1.1 $p_{i,j,k} \leftarrow p_{i,j,k} * 100 / d\left(p_{i,j,k}, p_{i_{selected\_probe}, j_{selected\_probe}, k_{selected\_probe}}\right)$
.
.
.

Figure 6.14: WCL-7 – seven choice algorithm – pseudocode.

Steps 3.3 to 3.12 are as in figure 6.12.

Again, the subject is able to audition C at any time; the interaction ends when the subject judges **C** and **T** to be indistinguishable. Data pertaining to the iteration is logged by the software; the degree to which the interaction has been successful is measured by the gradient of the approach of **C** to **T**, and the number of iterations required.

6.3.2.6.1. Rationale for multiplication factor value

As stated above, the values of each cells were multiplied by a factor inversely proportional to their distance from the cell corresponding to the chosen probe. This is a simple linear relationship, and is used because of this simplicity. Again, this will be revisited and discussed further in chapter eight.

6.4. Choice of platform

Before discussing the testing procedures, we briefly describe the development platform and the rationale for its use.

The programming language for the software was REALbasic for Mac OS X (REAL Software, 2006). This is a rapid application development environment which compiles native applications for Windows, Macintosh and Linux. It was chosen because of the speed with which software could be developed and tested. However, while it allows fast development, its provision for the generation and processing of audio is rudimentary. In order to incorporate audio into the test software, the programming language MacCsound, a dialect of the dedicated audio programming language Csound (Vercoe, 1985), was used to generate audio stimuli in the software described in this chapter; for the software described in the next chapter, PortAudio libraries (Bencina and Burk, 2004) were imported to generate the stimuli and target sounds.

Csound is a programming language, written in C, for audio synthesis, signal processing and sound design. Originally developed by Barry Vercoe at MIT, it is

free software issued under the GNU Lesser General Public License, and exists in versions for Windows, Mac OS 7, 8, 9 and X and Linux. Typically, a Csound program is contained in two text files; an 'orchestra' file (.orc), which defines the software generated 'instruments', and a score file which describes what those 'instruments' will play. (More recent versions, such as MacCsound combine these two components into one project file.) The Csound renderer then generates audio from these files, either in real time or as a sound file.

PortAudio is a cross-platform and open source audio library for the synthesis and processing of audio, which can be incorporated into programs written for Windows, Mac OS 8, 9 and X, and Unix. PortAudio RB, which is used here, is an implementation of PortAudio for REALbasic.

A full description of the program design is given in the appendix.

6.5. Procedure

This section describes the running and findings of a number of software tests conducted in the Sir John Cass Department of Art, Media and Design of London Metropolitan University between May 2nd and May 6th 2008, and November 7th and November 10th 2008.

Six versions of the software were prepared and were loaded onto six Apple Macintosh computers. These were as follows:

I: Multidimensional line search – formant space

II: Multidimensional line search – SCG-EHA space

III: WCL-2 – formant space

IV: WCL-2 – SCG-EHA space

V: WCL-7 - formant space

VI: WCL-7 – SCG-EHA space

The characteristics of the target sounds were as shown in figure 6.15.

	Target sound parameters	Initial Euclidean distance between weighted centroid and target
Formant space	Formant I centre frequency = 123.2 Hz Formant II centre frequency = 616 Hz Formant III centre frequency = 5447.119 Hz	8.124
SCG-EHA space	Attack time = 0.013 seconds EHA = 1 dB SCG = 6.938	6.403

Figure 6.15: Parameters of target sounds in the formant and SCH-EHA spaces.

These parameters place the target sounds near the edge of their respective attribute spaces. As the weighted centroid of the probability table will, at the outset, will be at its centre (because all the probability values are the same), this will facilitate the tracking of its movement.

6.5.1. Test procedures

Fifteen subjects were used for this test, who were paid for their time. The purpose of the test was explained, and each subject given a few minutes to practise operating the interfaces and to become accustomed to the sounds. Each subject was then asked to run each test **I** to **VI**; the order in which the tests were run varied randomly for each subject. Tests were conducted using headphones; in all cases, subjects were able to audition all sounds as many times as they wished

before making a decision.

6.5.1.1. Multidimensional line search (tests **I** and **II**)

Each subject was asked to manipulate the three software sliders, listening to the generated sound each time until EITHER the 'Play sound' button had been clicked on sixteen times OR a slider setting was found for which the generated sound was judged to be indistinguishable from the target. The choice of sixteen was pragmatically arrived at in the course of a number of pilot tests; it provided a sufficient search of the space for assessing the efficacy of the approach while minimising the risk of fatigue in the task. It was also noted that there was little or no further convergence on the target after about the sixteenth iteration, both in the pilot tests and in the results of the tests presented here. For this reason, sixteen iterations were chosen at the maximum number of iterations for all tests; this will be further discussed in section 6.7 and the conclusion of this chapter.

6.5.1.2. WCL-2 : two-alternative forced choice (tests **III** and **IV**)

Each subject was asked to listen to the target and then judge which of two sounds A or B more closely resembled it. After making the selection by clicking on the appropriate button, two new sounds A and B were generated by the software, and the process repeated until EITHER sixteen iterations had been completed OR the sound generated by clicking on the 'Candidate' button was judged to be indistinguishable from the target.

6.5.1.3. WCL-7: seven-alternative forced choice (tests **V** and **VI**)

Each subject was asked to listen to the target and then judge which one of

the seven sounds heard by clicking on the seven buttons labelled ''Listen to 1", "Listen to 2", etc more closely resembled it. After making the selection by clicking on the appropriate button, seven new sounds were generated by the software, and the process repeated until EITHER sixteen iterations had been completed OR the sound generated by clicking on the 'Candidate' button was judged to be indistinguishable from the target.

6.5.1.4. 'Control'

Finally, in order to determine whether the strategy was, in fact, operating in response to user input and was not simply generating spurious results, the WCL-2 strategy was run with a simulation of user input, where the 'user response' was entirely random.

6.6. Results

In this section, we will consider the results from the multidimensional line search, WCL-2 and WCL-7 strategies operating in the formant and SCG-EHA spaces, and compare them with the results of the 'control' version described in the previous section. We start with the results from the multidimensional line search, where subjects are supplied simply with sliders connecting to the axes of the space.

6.6.1. Multidimensional line search

Figure 6.16 shows the trajectory averaged over all fifteen subject interactions in the formant space. In all the graphs shown in this section, iteration

0 represents the default setting of the three sliders which were set at their mid-point at the beginning of the interaction.

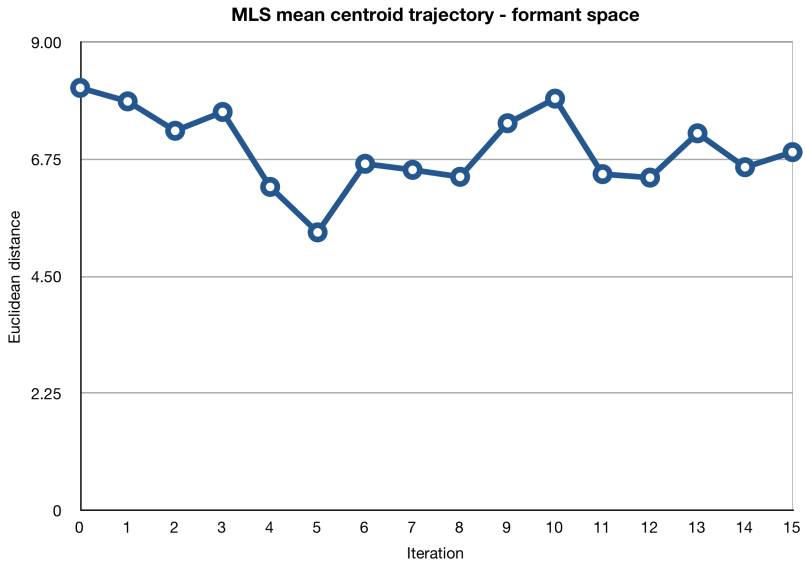


Figure 6.16: MLS mean trajectory of weighted centroid in formant space.

Overall, the first five iterations show a convergence on the target. Trajectories from iteration five onwards, however, become increasingly erratic, as subjects attempted to 'fine tune' the sound arrived at.

Many subjects began the search by adjusting all three sliders to their minimum value (this accounts for the considerable change in value seen between iterations 0 and 1 in figures 6.17 (a), (b) and (c), and incrementally adjusted the value of a single slider for a few iterations before turning their attention to another one.

Also of interest is the trajectory projected along each axis of the formant space, from iteration 1 onwards, shown in figures 6.17 (a), (b) and (c) . Only along the formant III axis is there evidence of subsequent convergence on the target. This is somewhat at variance with the results of the listening tests described in chapter

five, where it appeared that the formant II axis was most salient in the perception of timbral distance in this space.

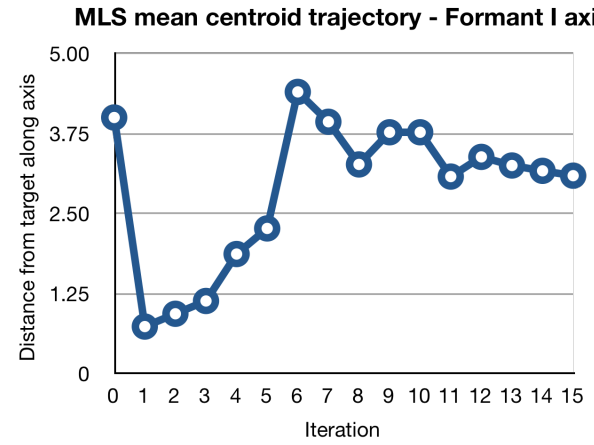


Figure 6.17(a): Trajectory of weighted centroid in formant space projected on formant I axis

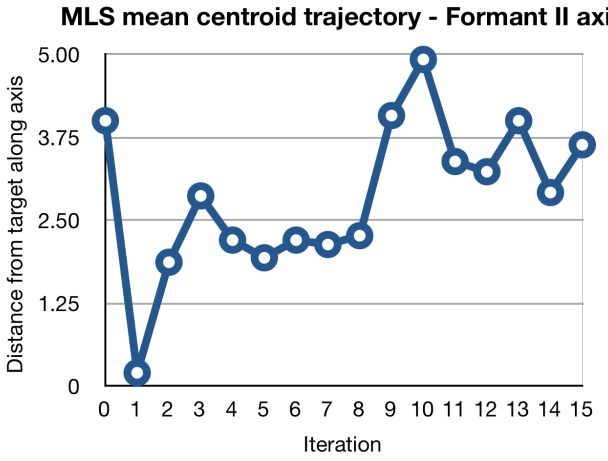


Figure 6.17(b): Trajectory of weighted centroid in formant space projected on formant II axis

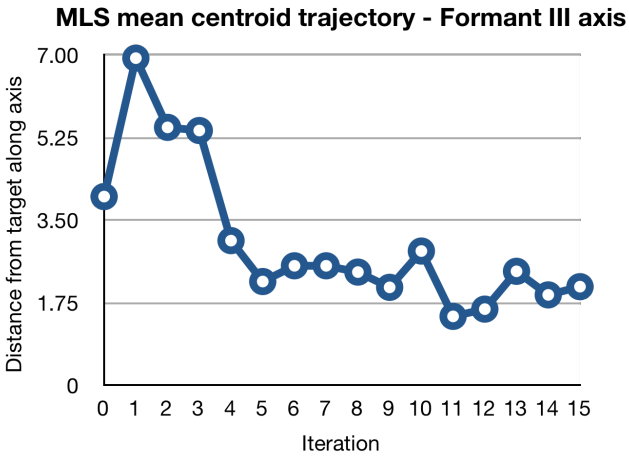


Figure 6.17(c): Trajectory of weighted centroid in formant space projected on formant III axis

Turning now to the operation of the strategy in the SCG-EHA space, defined in section 6.2.2.3, we see in figure 6.18 a much clearer overall convergence. Again, many subjects began the search by moving the sliders to their minimum value, thus causing the overall jump in value between iterations 0 and 1. After iteration 1, the individual trajectories show, in most cases, the probe sound approaching the target.

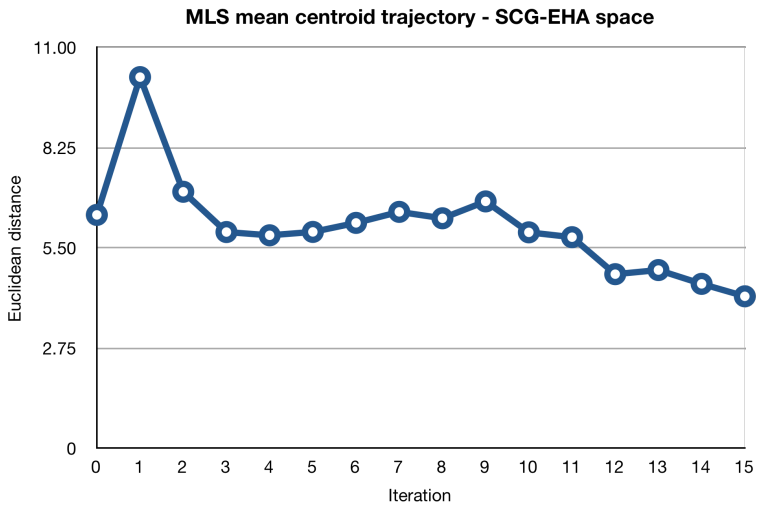


Figure 6.18: MLS mean trajectory of weighted centroid in SCG-EHA space.

The projection of the above curve on the three axes is shown in figures 6.19(a), (b) and (c).

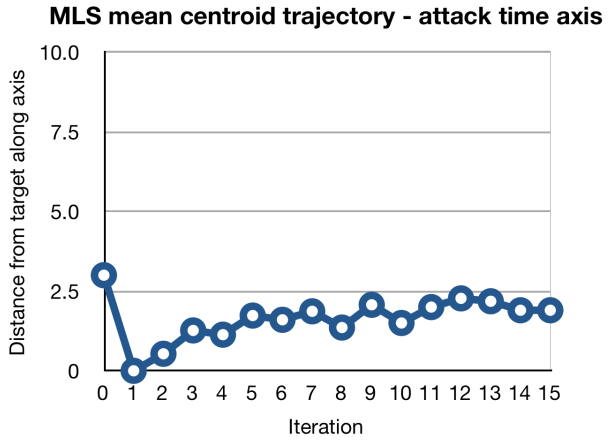


Figure 6.19(a): Trajectory of weighted centroid in SCG-EHA space projected on attack time axis.

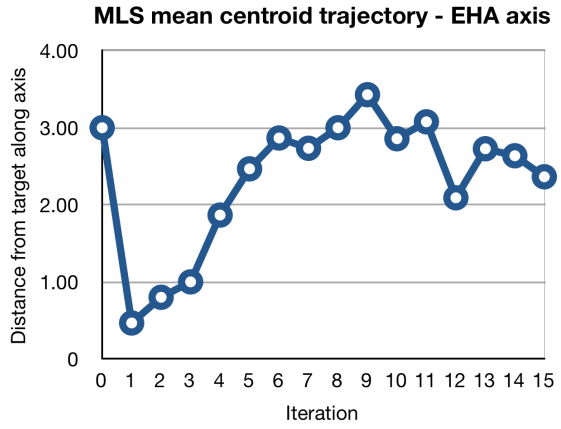


Figure 6.19(b): Trajectory of weighted centroid in SCG-EHA space projected on the EHA axis.

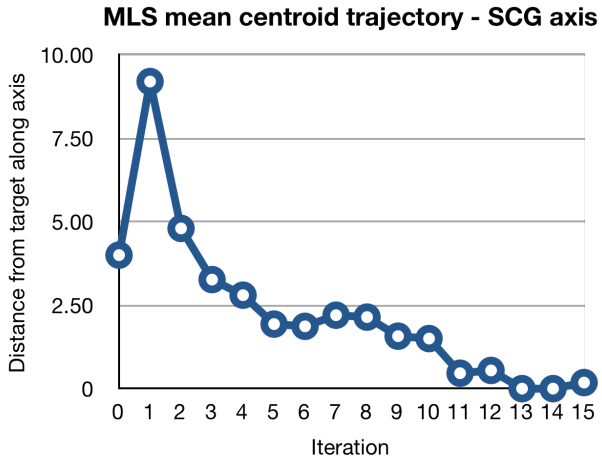


Figure 6.19(c): Trajectory of weighted centroid in SCG-EHA space projected on the SCG axis.

Here, we see that the approach to the target along the SCG axis (the attribute associated with brightness) is fairly smooth; the other two trajectories, by contrast, move away from the target. This is particularly curious in the case of the attack time axis; attack time is a very audible and salient characteristic of sound, and the expectation was that there would have been a swift convergence on the target along this axis. Inspection of the individual trajectories showed, however, that most of them ended on a value that was, in fact, close to the target. The apparent drift away from the target, shown in figure 6.19(a) appears to have two reasons. Firstly, as was seen in the equivalent experiment in the formant space, many subjects began the search by setting all three sliders to their minimum value. This was, in fact, very close to the target value (see iteration 1 in figure 6.19(a)); the

subsequent drift away can be explained by subjects attempting to 'fine tune' this parameter. Secondly, the results were distorted by three atypical trajectories whose overall movement was away from the target.

In both spaces, it appears that one attribute of the sound is preferred and used as a guide for the task. The use of the SCG (brightness) parameter in the SCG-EHA in this respect is entirely congruent with the findings of a number of timbre perception studies, reviewed in chapter three.

6.6.2. WCL-2: two-alternative forced choice

We turn now to consider the data from the two-alternative forced choice tests. Figure 6.20(a) shows the change in the Euclidean distance in the formant space between the weighted centroid of the probability table and the cell in the probability table corresponding to the target, for all fifteen subjects. Nearly all of the fifteen traces show a reduction in the distance. Figure 6.20(b) shows the average trajectory followed by the weighted centroid relative to the target - this is calculated by taking the mean distance at each iteration point for all fifteen subjects.

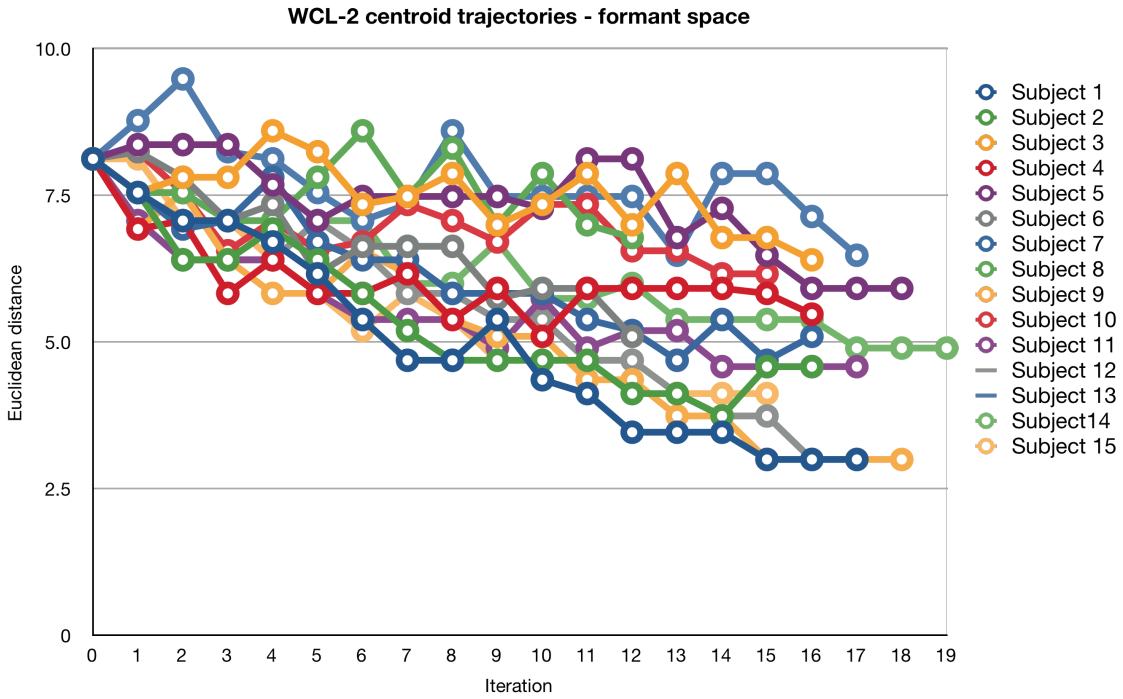


Figure 6.20(a): Weighted centroid trajectories in formant space using WCL-2 strategy.

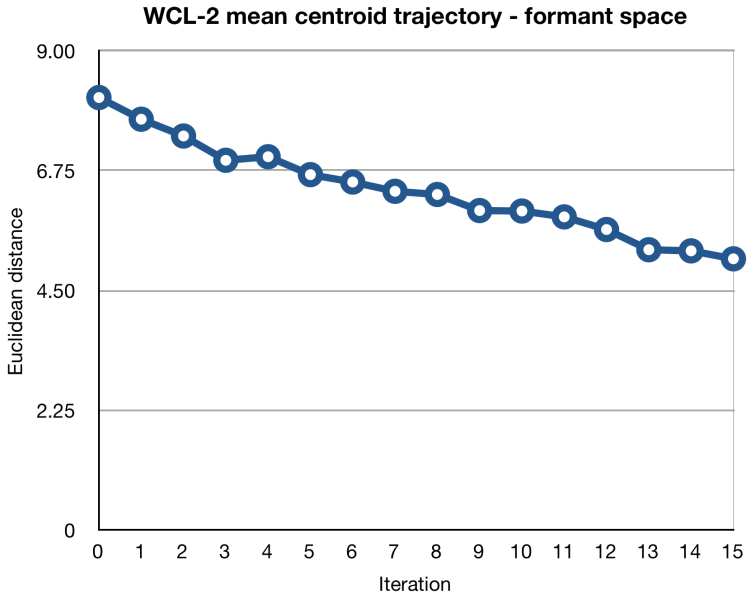


Figure 6.20(b): Mean weighted centroid trajectory in formant space using WCL-2 strategy.

Overall, there is a steady reduction in the target/weighted centroid distance from 7.72 at the first iteration to 5.1 at the fifteenth.

Bearing in mind that the success of this strategy was entirely dependent on the subject being able to 'correctly' identify which of the two probes was closer to the target, it is worth noting that the mean percentage of 'correct' identifications

was 73.74% - entirely consistent with the results of the listening tests described in chapter five.

Turning now to consider the data from the SCG-EHA space, we again see a reduction in the target/weighted centroid distance for most subjects.

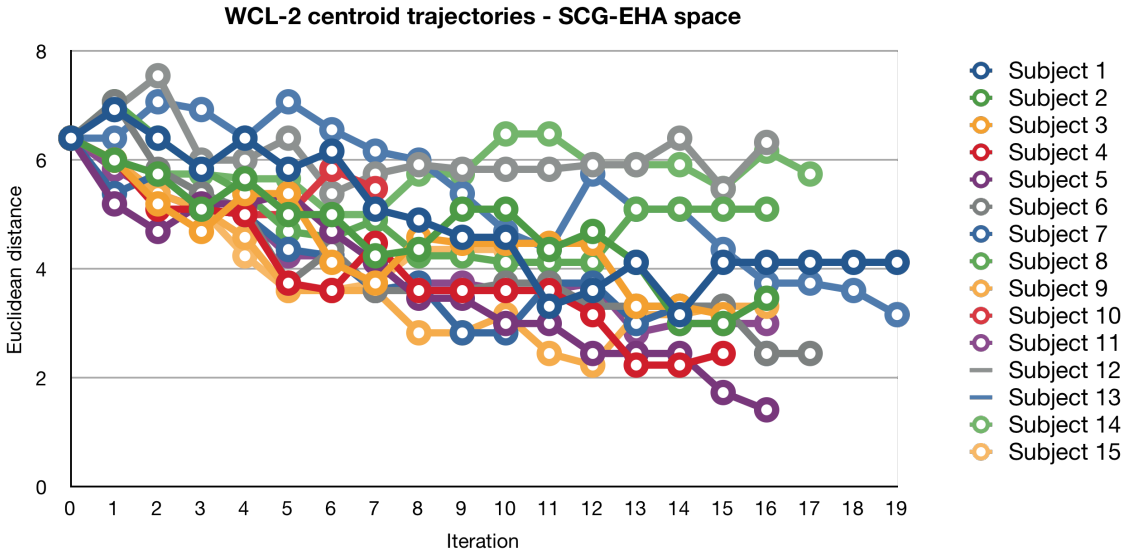


Figure 6.21(a): Weighted centroid trajectories in SCG-EHA space using WCL-2 strategy.

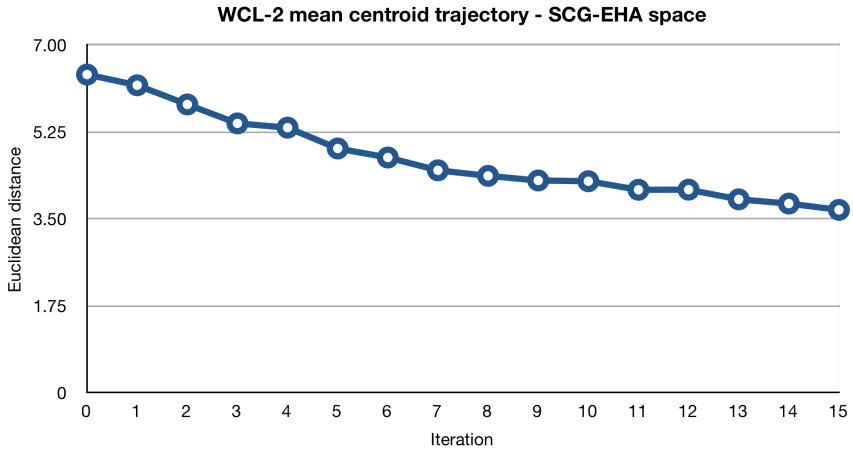


Figure 6.21(b): Mean weighted centroid trajectory in SCG-EHA space using WCL-2 strategy.

The mean trajectory (figure 6.21(b)), calculated as before, shows a reduction from 6.19 at the first iteration to 3.68 at the fifteenth.

The mean percentage of 'correct' identifications in this attribute space was 83.3% - rather higher than in the formant space. Nevertheless, there is a steeper gradient in the mean trajectory in the formant space than in the SCG-EHA space.

6.6.3. WCL-7: seven-alternative forced choice

The data from the seven-alternative forced choice version is discussed here. Again, we see an overall reduction in distance in nearly all of the fifteen traces (figure 6.22(a)), a trend which is encapsulated in the mean trajectory shown in figure 6.22(b).

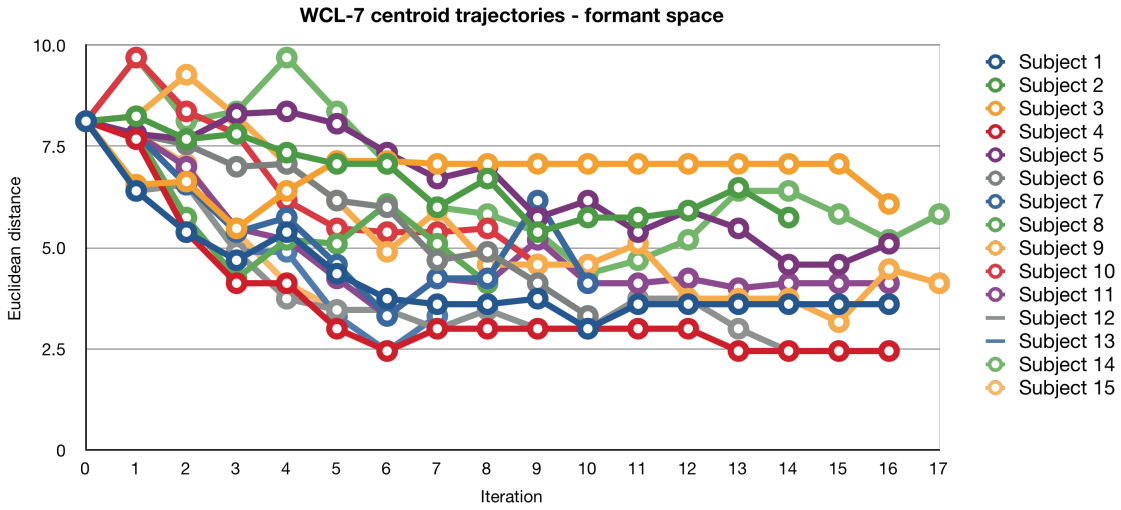


Figure 6.22(a): Weighted centroid trajectories in formant space using WCL-7 strategy.

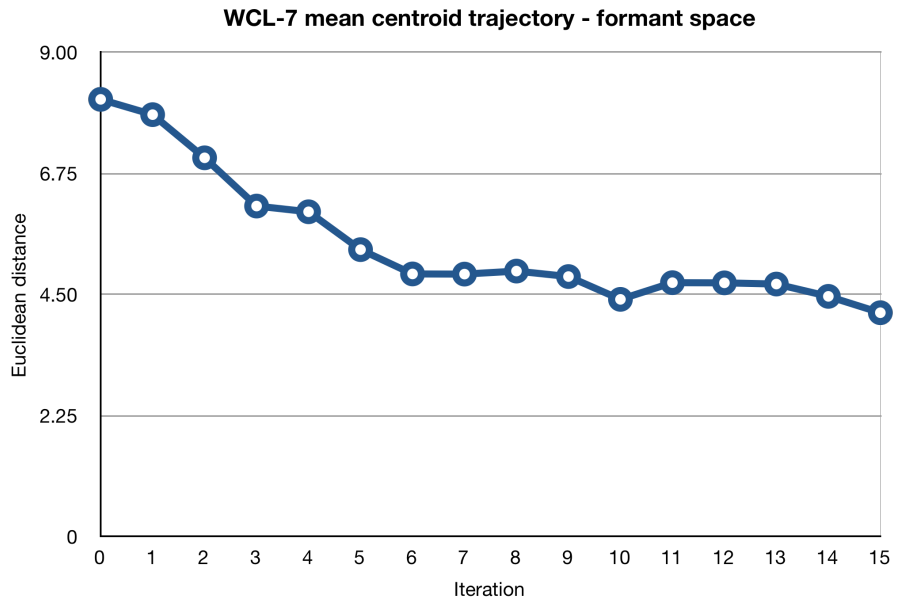


Figure 6.22(b): Mean weighted centroid trajectory in formant space using WCL-7 strategy.

Overall, there is a steady reduction in the target/weighted centroid distance from 7.84 at the first iteration to 4.98 at the fifteenth.

Figures 6.23(a) and 6.23(b) show the equivalent trajectories for the SCG-EHA space.

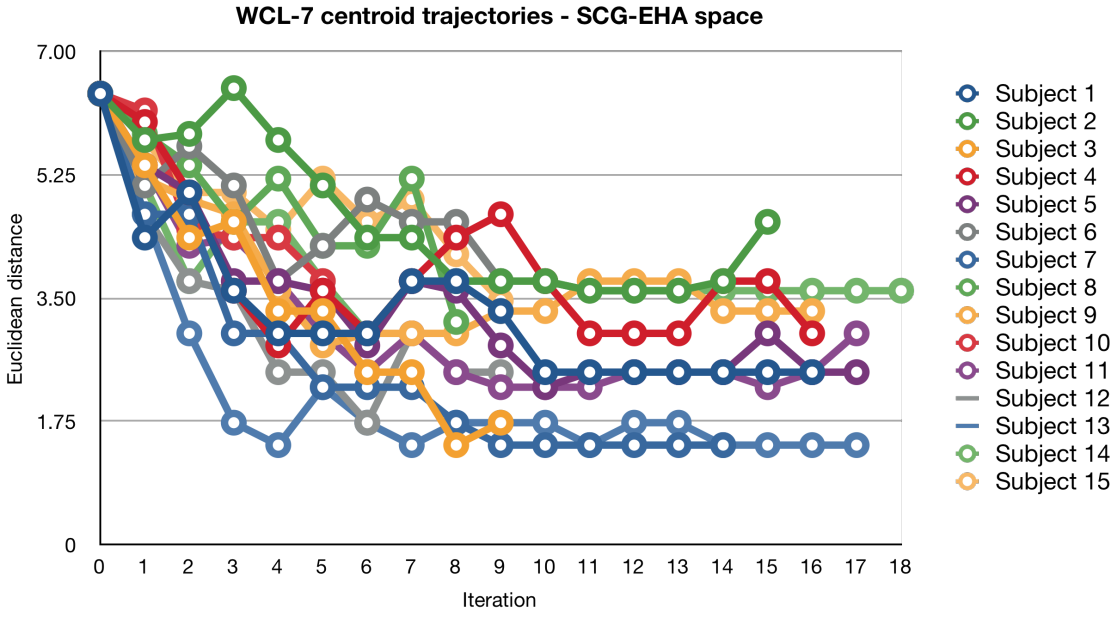


Figure 6.23(a): Weighted centroid trajectories in SCG-EHA space using WCL-7 strategy.

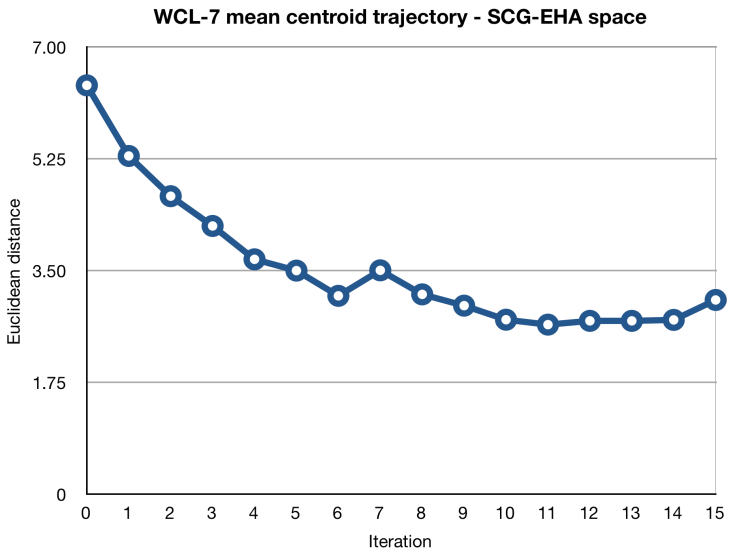


Figure 6.23(b): Mean weighted centroid trajectory in SCG-EHA space using WCL-7 strategy.

Here we see a reduction in target/weighted centroid distance from 5.29 at the first iteration to 2.9 at the seventeenth. There is a noticeable spike in the trajectory in figure 6.23(b) around the seventh iteration, which can also be seen in a number of individual trajectories in figure 6.23(a). It was thought that this could be attributable to the pseudo-random generation of probe coordinates, in which the same sequence of probe sets is generated for all subjects. Inspection of the control data for all the tests, however, revealed this not to be the case – in fact, different sequences of probes were generated for each subject. No further explanation is offered for this.

One notable statistic which emerged from the seven-alternative forced choice data is the mean percentage of 'correct' identifications – 48.4% in the formant space and 67.56% in the SCG-EHA space. Given that subjects were asked to choose one from seven rather than one from two, this is remarkable.

6.6.4. 'Control' results

Figures 6.24(a) to 6.24(d) show the weighted centroid trajectories when random 'user' responses were given in the WCL-2 and WCL-7 versions of the search strategy. Both of them show the characteristics of Brownian motion, or the 'drunkard's walk' – the random path taken by, for example, a particle suspended in a fluid. The mean trajectory in both cases was more or less a straight line. In the case of the 'random' WCL-2 strategy, because the user responses was randomly generated, they were very often 'correct'; the percentage of correct responses varied considerably in each of the ten 'random' runs which were executed. This can be seen in figure 6.24(a); those runs for which the percentage of 'correct' responses were 50% and above show downward (i.e. convergent) trajectories, with the steepest gradient resulting from a 'correct' response percentage of 89.47%. Conversely, those runs where the percentage of correct responses was less than 50% showed upward trajectories.

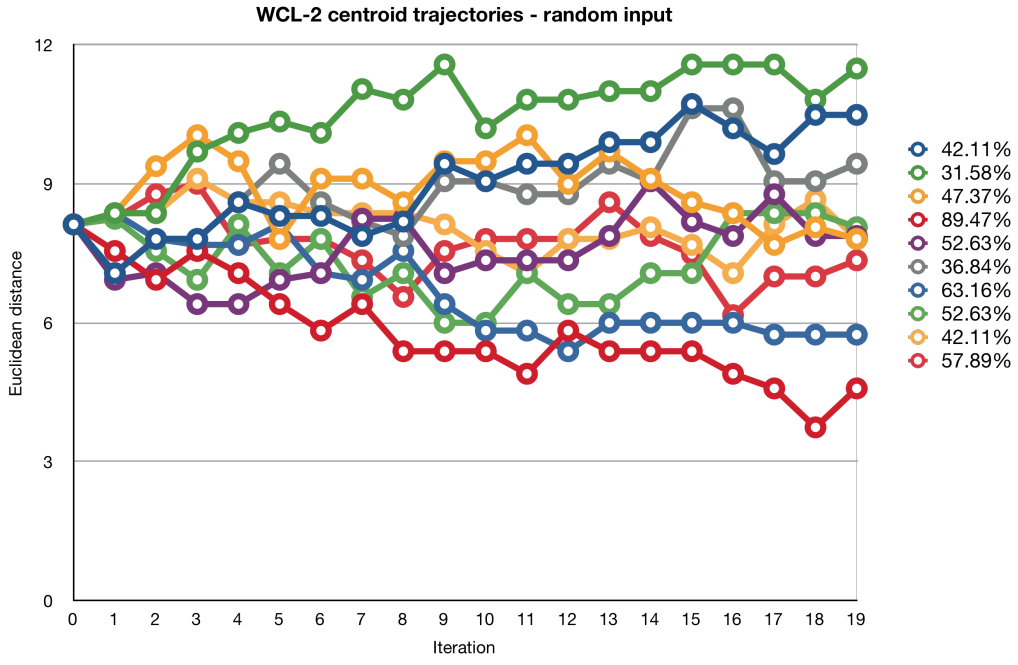


Figure 6.24(a): Weighted centroid trajectories using random user input for WCL-2 strategy.

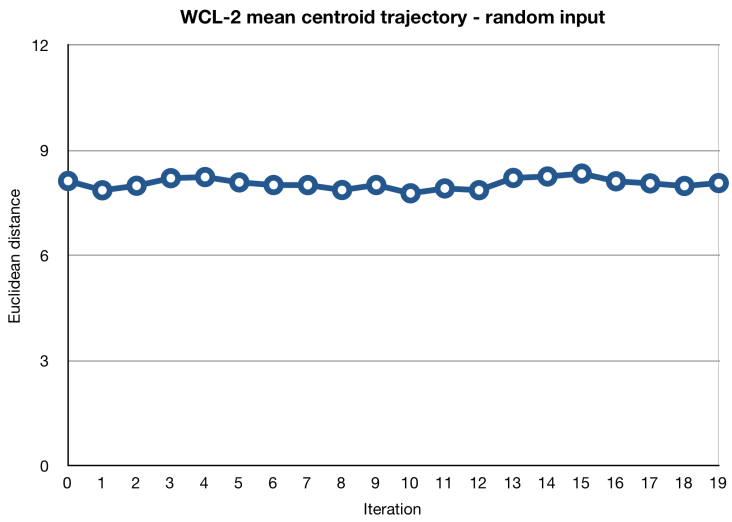


Figure 6.24(b): Mean weighted centroid trajectory using random user input for WCL-2 strategy.

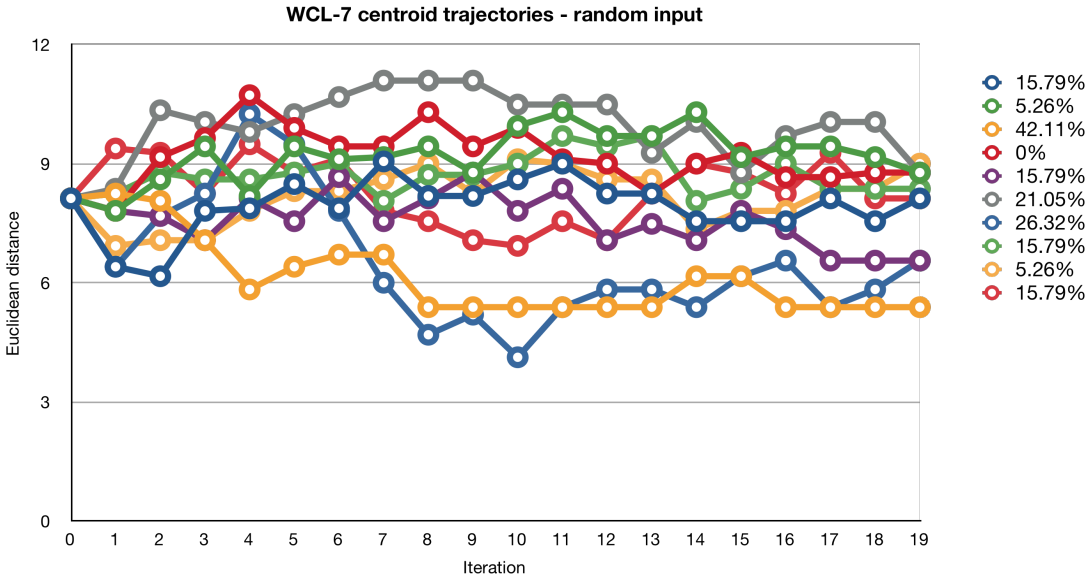


Figure 6.24(c): Weighted centroid trajectories using random user input for WCL-7 strategy.

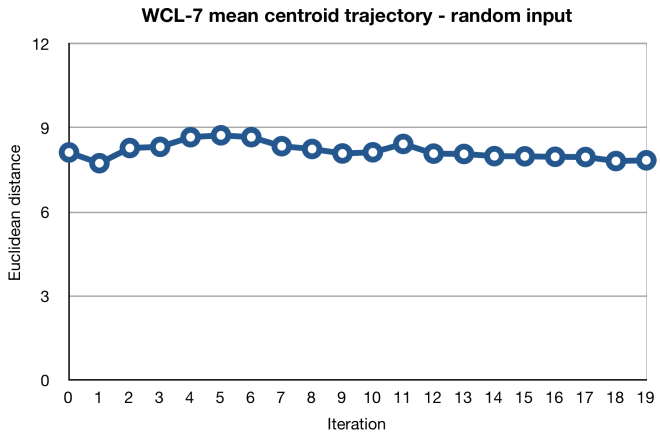


Figure 6.24(d): Mean weighted centroid trajectory using random user input for WCL-7 strategy.

6.7. Summary and discussion of results

Figures 6.25(a) and 6.25(b) summarise the mean WCL-2 and WCL-7 weighted centroid trajectories in the formant and SCG-EHA attribute spaces respectively; in each case, they are compared with the trajectory in the respective spaces of the sound generated by the user on each iteration of the multidimensional line search strategy (again, averaged for all fifteen interactions)

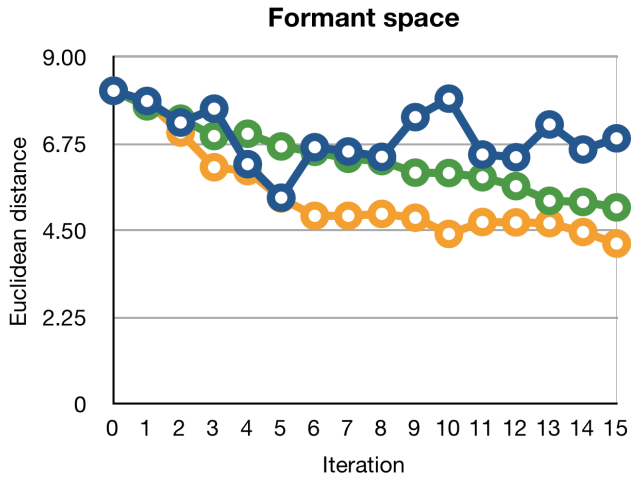


Figure 6.25(a): Weighted centroid trajectories in formant space using WCL-7 strategy.

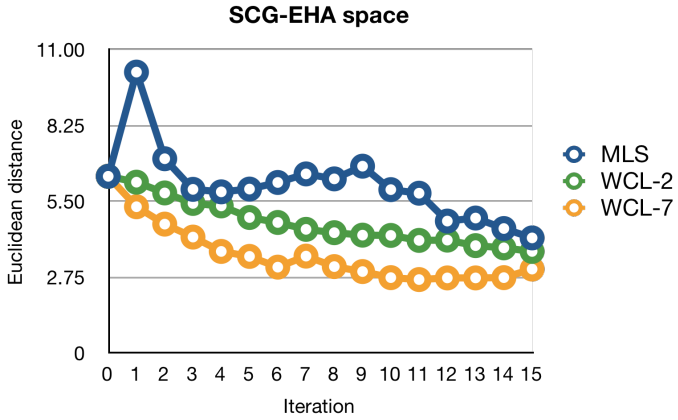


Figure 6.25(b): Summary of mean weighted centroid trajectories in a) formant space and b) SCG-EHA space.

In all three strategies deployed in the two attribute spaces, there was considerable variation in individual subject performance. However, the mean trajectories of the WCL-2 and WCL-7 strategies show a greater gradient (faster convergence on the target) than that of the MLS strategy, with the WCL-7 trajectory being, in both cases, the steepest.

As stated in section 6.5.1.1, subjects were limited to a maximum of sixteen iterations. However, a number of subjects did not terminate the interaction after the sixteenth iteration, but continued working with the interface. Inspection of the the trajectories showed that they either flattened out or became increasingly erratic after about sixteen iterations; this can be most easily seen in the example individual trajectory shown in figure 6.26.

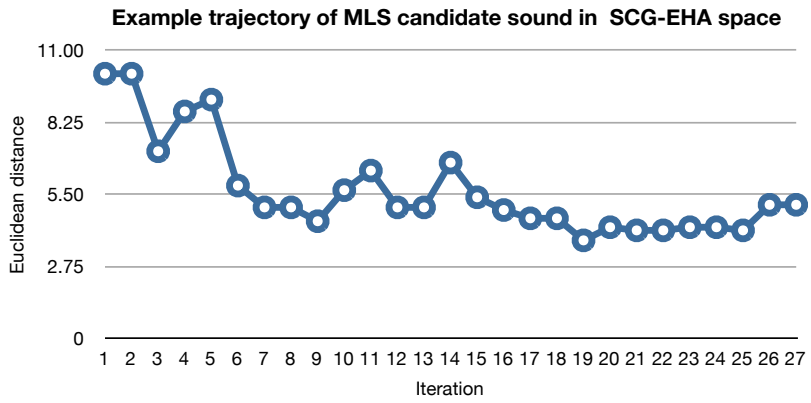


Figure 6.26: Trajectory through SCG-EHA space taken by one subject.

Further discussion of these findings is deferred to the next chapter, where they will be considered alongside the results of MLS, WCL-2 and WCL-7 tests in a seven-dimensional space.

Chapter 7 - Searching a multidimensional MDS space

7.1. Introduction

The two attribute spaces examined in the previous chapter are very simple and limited in their coverage. In order to determine how effectively the WCL method might operate in a more realistic and 'real world' attribute space, we now need to test it against a wider and more musically useful range of timbres. To build such a space, we can begin by assembling a palette of sounds drawn from a list of orchestral musical instruments whose timbres are very diverse. As was discussed in chapter three, such an attribute space would necessarily be highly multidimensional; chapter five, however, showed how data reduction techniques such as MDS and PCA can be used to represent these sounds in a space of reduced dimensionality while, at the same time, preserving most of the variance between them. It is claimed in this chapter that such a space can be used as a vehicle for synthesis.

This chapter outlines why the assumption of the tractability of such a step is justifiable, discusses the method by which the attribute space to be used in the present study is constructed, reports the testing procedure and finally discusses the results.

7.2. Multidimensional scaling (MDS) – rationale for its use

MDS has been shown to be effective both as a means of determining salient features of timbre and for representing similarity/dissimilarity relationships

between timbres in an attribute space of reduced dimensionality. Whether such a space could be used as a synthesis space, however, has been questioned (McAdams, Beauchamp *et al.*, 1999). The authors point out that it is not obvious that the identified components each correspond to a clearly identified acoustical quantity that could be varied in sound synthesis. This, in turn, implies that simply providing a user with a set of sliders each of which corresponds to a principal component (which is the essence of the MLS method) might not be a successful strategy. However, three findings that emerge from the work of Hourdin *et al* , reviewed in chapter four, do seem significant.

Firstly, particular trajectory curves in the derived space seemed to be associated with particular timbres. While no listening tests were performed to verify this, the authors suggest that this may be 'an interesting tool for composers' – certainly it implies a degree of congruence between the physical and perceptual spaces. Figure 7.1 shows the trajectories in the reduced space of a) a muted trombone and a tenor trombone, b) a marimba and xylophone and c) a guitar and archlute. The instruments in each pair show very obvious similarities in their paths through the space.

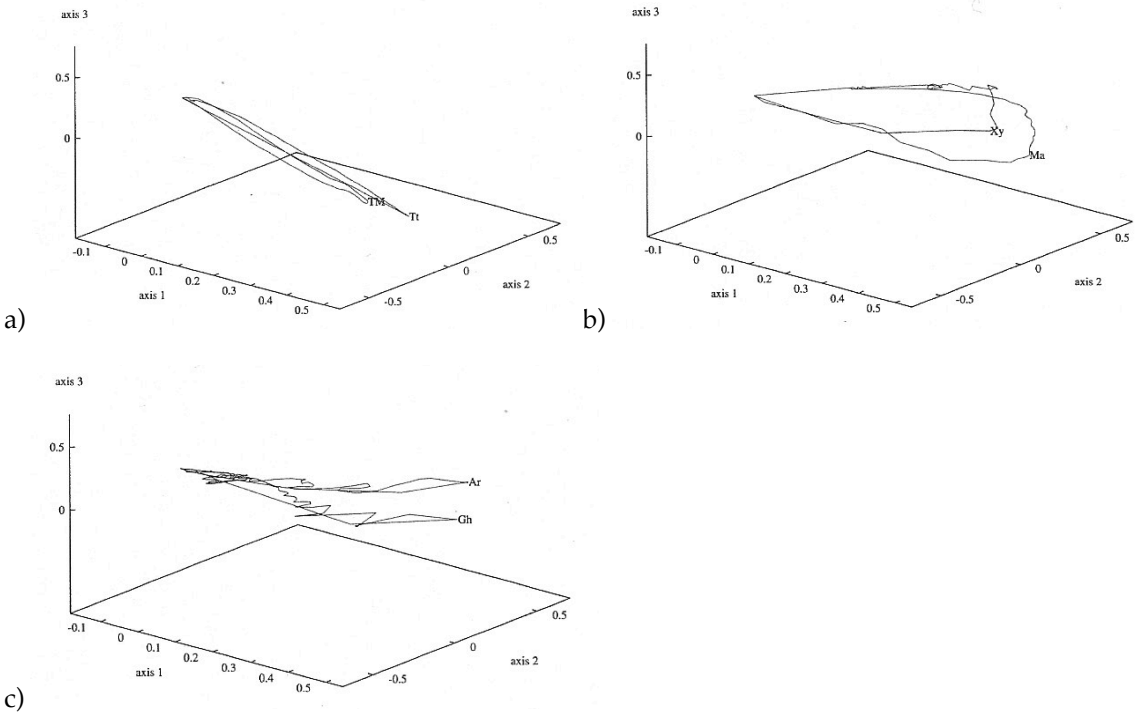


Figure 7.1: Representation in the reduced space of similar timbres: (a) trombone muted TM and tenor trombone Tt (b) marimba Ma and xylophone Xy (c) guitar Gh and archlute Ar - from Hourdin *et al* (1997).

Secondly, the space seems to be stable and predictive – when intermediate curves were plotted which were interpolations between two existing curves (in this case, tenor trombone and cello played *martelé*), the resultant tone sounded plausibly like a hybrid of these two (Hourdin, Charbonneau *et al*., 1997). Thirdly, the reconstructed tones were musically acceptable, even where the number of factors was relatively low.

For these reasons, a MDS space, drawing on the work of Hourdin *et al*, was used in order to test the MLS and WCL search strategies.

7.3. Derivation and construction of the attribute space

We briefly summarise the MDS process here, before going on to consider in greater detail its implementation in the search software.

As discussed in chapter four, the input to an MDS analysis is typically a matrix of 'proximities' (which may represent similarity/dissimilarity judgments, for example) between a set of objects. The output is a geometric configuration of points, each representing a single object in the set, such that their disposition in the space approximates their proximity relationships. Each dimension of the space is called a 'factor'; the first one captures the maximum variance of the data, the second captures the second highest amount of variance and so on.

More specifically; a square and symmetric matrix **D** which we will call a *distance matrix,* is input to the process. A matrix cell $\mathbf{d}_{i,j}$ represents the magnitude of the relationship (proximity, similarity etc) of items **i** and **j**. MDS then seeks a configuration of these items in a space **C** with a specified number of dimensions, such that the Euclidean distances between the items correspond as closely as possible to the distance matrix. The *stress* measure is typically used to evaluate how well or poorly this has been done - the smaller the stress value, the better the fit.

Where the intention is to represent a dataset in a space of lower dimensionality than that of **C,** we can make use of the vector **E** of eigenvalues of the scalar product matrix **CC'** (where **C'** is the transpose of **C**). If the eigenvalues of the first **k** elements of **E** are significantly greater than the remainder of them, we can use the first **k** columns (i.e. dimensions) of **C** to construct a space of reduced dimensionality.

The attribute space to be constructed is seven dimensional. Six dimensions are derived by MDS techniques. The seventh is attack time, with the same characteristics as those of the SCG-EHA space described in the previous chapter.

The choice of a six dimensional space will be justified in the section describing the MDS process in detail.

As stated previously, both the space and the construction technique used to build it are derived in part from the work of Hourdin *et al*; the list of fifteen instrumental timbres is broadly the same as that used in that study, and is as follows.

1. Alto flute
2. Alto saxophone (no vibrato, *fortissimo*)
3. Alto saxophone (no vibrato, *mezzo forte*)
4. Bass clarinet
5. Bass flute
6. Bassoon
7. Bb trumpet
8. Cello (*sul* A)
9. Eb clarinet
10. Flute
11. French horn
12. Oboe
13. Soprano saxophone
14. Tenor trombone
15. Viola (*sul* G)

The samples were taken from the sample library of the University of Iowa Electronic Music Studios (Fritts, 1997); all samples were recorded anechoically in mono, 16 bit, 44.1 kHz AIFF format. The pitch of all the instrumental sounds was Eb above middle C (311 Hz); and all were played *mezzo forte,* except where otherwise indicated. Each instrumental sample was then edited to remove the

onset and decay transients, leaving only the steady state portion, which was, in all cases, 0.3 seconds.

The following section describes the process by which the reduced space, and in particular the number of its dimensions, was arrived at.

7.3.1. Derivation

The first stage is to establish the number of dimensions required to represent the audio samples with minimum loss of information. Clearly, the fewer dimensions used, the greater the loss of data is likely to be. However, a higher dimensionality brings with it a greater computational cost; the addition of one extra dimension increases the amount of data to be processed (in this particular system) by a factor of seven. There is, therefore, a trade off between the accuracy of the data recovery and system response times. It will be shown in the course of the following discussion that the use of six MDS dimensions results in an acceptable level of data loss.

The process is summarised in figure 7.2, and discussed in the following sections.

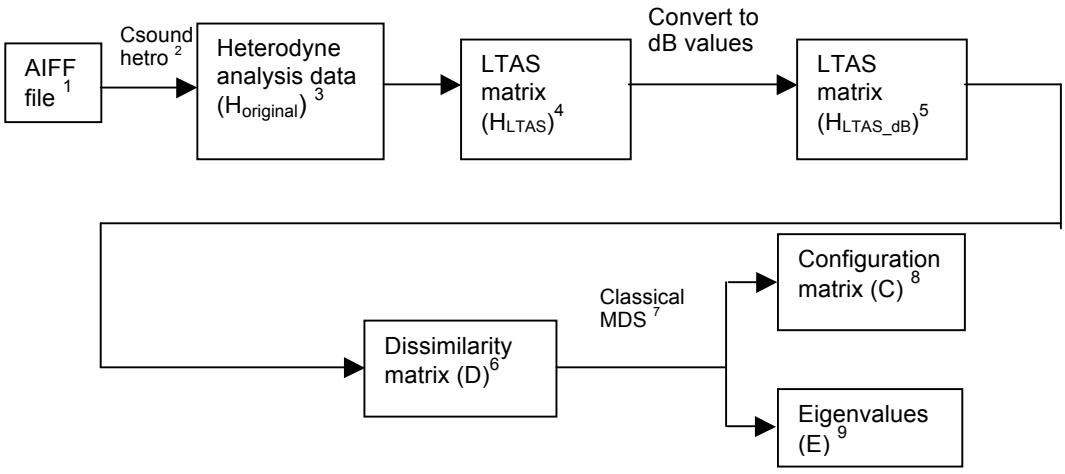


Figure 7.2: Multidimensional scaling of instrument samples.

1. All the edited samples were normalized to –3 dB relative to full amplitude, using sound editing software and spliced together to form one single AIFF file.

2. The audio file was processed using *heterodyne filter analysis*, discussed in chapter four in connection with the work of Hourdin *et al* . To summarise: heterodyne filtering resolves periodic or quasi-periodic signals into component harmonics, given an initial fundamental frequency: the multiplication of the input waveform by a sine and cosine function at harmonic frequencies and the summing of the results over a short time period yields amplitude and phase data for each harmonic. The implementation used here was that provided as a utility (*hetro*) as part of the Csound audio programming environment (Klapper, 2000).

3. The output of *hetro* is a matrix of data in which the columns contain the time-varying amplitude and frequency values of each harmonic[20], and each row is a breakpoint snapshot of the instantaneous spectrum, as shown in figure 7.3 for an analysis consisting of *n* harmonics, and N rows[21].

[20] Phase information is not generated by *hetro* and is, in any case, not required here.

[21] The data output from the *hetro* utility is typically used for resynthesis, and is partially in binary format. In order to allow editing of the data and to re-format it for further processing in MATLAB, a utility program was written which took the *hetro* data as input and converted it into a comma-separated file.)

time	1st harm.		2nd harm.		3rd harm.		.		*n* th harm.	
t_0	$A_{0,0}$	$F_{0,0}$	$A_{1,0}$	$F_{1,0}$	$A_{2,0}$	$F_{2,0}$	.	.	$A_{n-1,0}$	$F_{n-1,0}$
t_1	$A_{0,1}$	$F_{0,1}$	$A_{1,1}$	$F_{1,1}$	$A_{2,1}$	$F_{2,1}$	.	.	$A_{n-1,1}$	$F_{n-1,1}$
t_2	$A_{0,2}$	$F_{0,2}$	$A_{1,2}$	$F_{1,2}$	$A_{2,2}$	$F_{2,2}$	.	.	$A_{n-1,2}$	$F_{n-1,2}$
t_3	$A_{0,3}$	$F_{0,3}$	$A_{1,3}$	$F_{1,3}$	$A_{2,3}$	$F_{2,3}$	.	.	$A_{n-1,3}$	$F_{n-1,3}$
.	.	.	.	.	.	.	.	.	.	.
t_{N-1}	$A_{0,N-1}$	$F_{0,N-1}$	$A_{1,N-1}$	$F_{1,N-1}$	$A_{2,N-1}$	$F_{2,N-1}$	.	.	$A_{n-1,N-}$	$F_{n-1,N-1}$

Figure 7.3: Output of the heterodyne filter process.

Because we are concerned with steady state spectra, the columns representing harmonic frequency fluctuations (F) were not included in the analysis and were removed – thus the N x 40 matrix becomes an N x 20 one.

4. A new 15 x 20 matrix H_{LTAS} was generated from the heterodyne data matrix $H_{original}$, such that each row holds the Long Time Averaged Spectrum (LTAS) for one instrumental sound.

5. The heterodyne data contains linear harmonic amplitudes. These should be converted to decibels, firstly, to be consistent with the space and secondly, because logarithmic rather than linear axes more closely align with amplitude perception. This is done as follows:

$$H_{LTAS\_dB} = 20\log(H_{LTAS})$$

6. A dissimilarity matrix D was built from H_{LTAS_dB} using the *pdist()* function in MATLAB (The Mathworks, 2007) . This is a 15 x 15 matrix whose (ij)th

element is equal to the Euclidean distance between the (i)th and (j)th points in H_{LTAS}.

7. The dissimilarity matrix D was used as input to a classical multidimensional scaling function *cmdscale().* This has two outputs, described below.

8. The first output is a 15 x p configuration matrix C , where p < 15, which is a solution space to the input dissimilarity matrix D (and which may or may not be identical to H_{LTAS_dB}).

9. The second output is a vector E holding the eigenvalues of C*C′. These are as shown in figure 7.4.

```
eigvals =

. 1.0e+04 *

. 4.4933
. 1.0093
. 0.5864
. 0.3160
. 0.1462
. 0.1442
. 0.1113
. 0.0854
. 0.0556
. 0.0403
. 0.0306
. 0.0150
. 0.0086
. 0.0052
. 0.0000
```

Figure 7.4: Eigenvalues of C*C′.

Each eigenvalue corresponds to one axis of the p-dimensional configuration matrix C; its magnitude indicates the relative contribution of the corresponding axis to the building of the dissimilarity matrix D (in other words, the amount of information associated with that axis). We can express the above eigenvalues as percentages of the total amount of information – see figure 7.5.

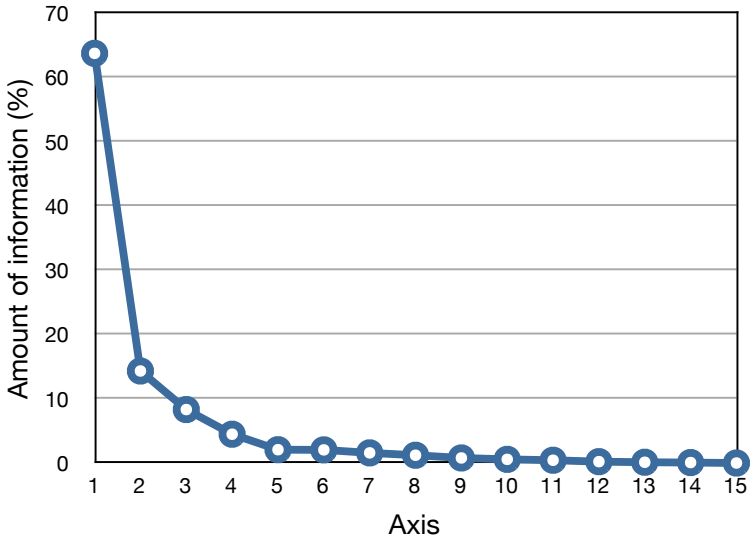


Figure 7.5: Eigenvalues of C*C′.

Note that the first six eigenvalues are considerably greater in magnitude than the remaining nine. Recasting the data as the cumulative percentage of information associated with an increasing number of eigenvectors (i.e. axes) (figure 7.6), we can see that 95 % of the total information required to reconstruct the spectra is associated with just six axes; thus, MDS can be used to reduce the dimensionality from twenty to six with minimal loss of information.

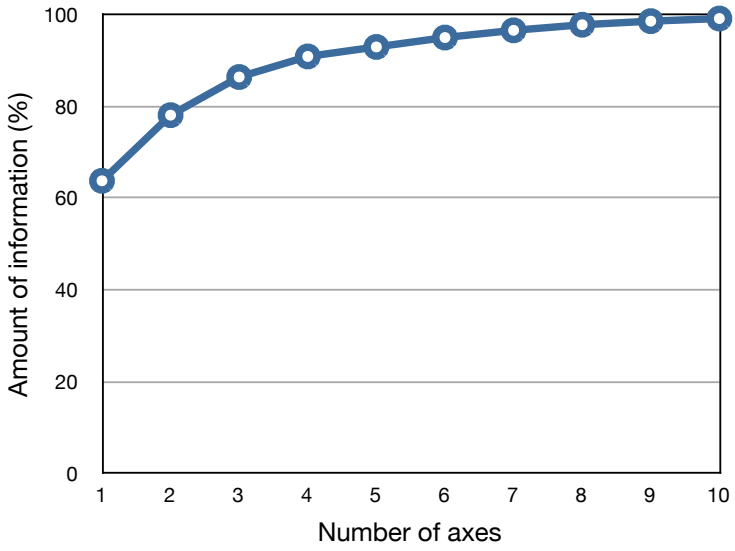


Figure 7.6: Eigenvalues of C*C′ as percentage of total information

7.3.2. Construction of the reduced dimensionality space

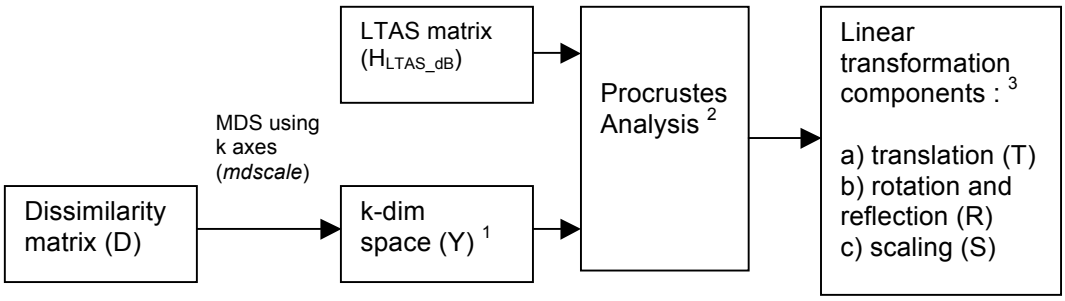


Figure 7.7: Process of building the reduced dimensionality space.

1. A new 6-dimensional space **Y** was generated using MDS from the dissimilarity matrix **D**. This was done in MATLAB using the *mdscale()* function. The features of this space are discussed further in sections 7.3.3 to 7.3.5.

2. The alignment of points in the reduced space is such that *translation* (centering the data), *rotation*, *reflection* and *scaling* are needed in order to recover, with minimum error, the original heterodyne data for synthesis. The data to do this was obtained using the *procrustes()* function in MATLAB; this is a function which determines a linear transformation of the points in one matrix which best conforms them to those in another. In this case, the two matrices are the reduced space just generated and the original matrix holding the Long Time Averaged Spectrum (LTAS) for each instrumental sound (H_{LTAS_dB}).

7.3.3. The reduced dimensionality space

The following scatter graph (figure 7.8) shows the fifteen instrumental sounds placed in a three dimensional space (the first three columns of the reduced space dataset).

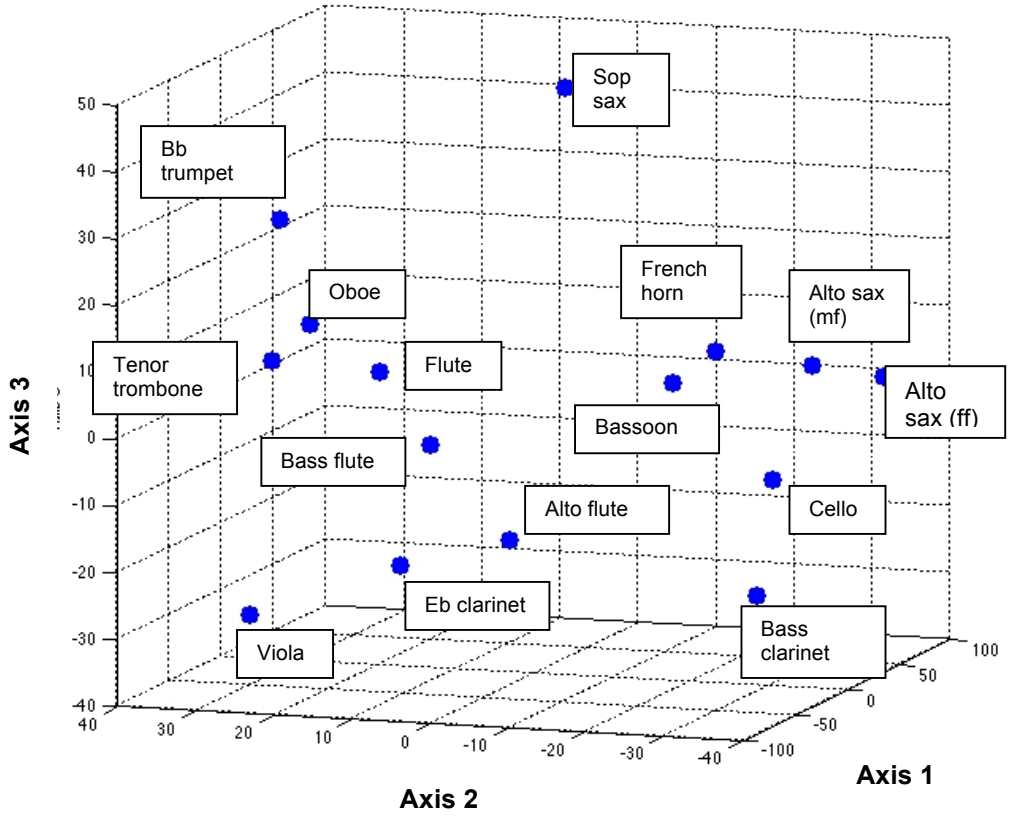


Figure 7.8: The 15 instrumental sounds located in a three dimensional space following MDS analysis.

7.3.4. Stability of the space

In order to verify the stability of this space, two instruments – alto flute and trumpet - were removed from the list, leaving thirteen instruments remaining, and the space reconstructed. (A similar test, described in section 4.3.2.2. of chapter four, was performed by Hourdin *et al* (1997) to verify the MDS space that they used).

Figures 7.9 (a) (b) and (c) show the two spaces generated, projected on axes 1 and 2, axes 1 and 3 and axes 2 and 3. The left hand column shows the space generated using the original fifteen instruments; the right hand column shows that generated using the list with the two instruments - alto flute and trumpet - removed.

Tenor trombone	Trmb
Bb trumpet	Trpt
Viola	Vl
Eb clarinet	Clt
Oboe	Ob
Flute	Fl
Bass flute	BFl
Alto flute	AFl

Sop saxophone	SSx
Bass clarinet	BClt
Cello	Cello
Alto sax (ff)	ASx-ff
Alto sax (mf)	ASx-mf
Bassoon	Bn
French horn	FH

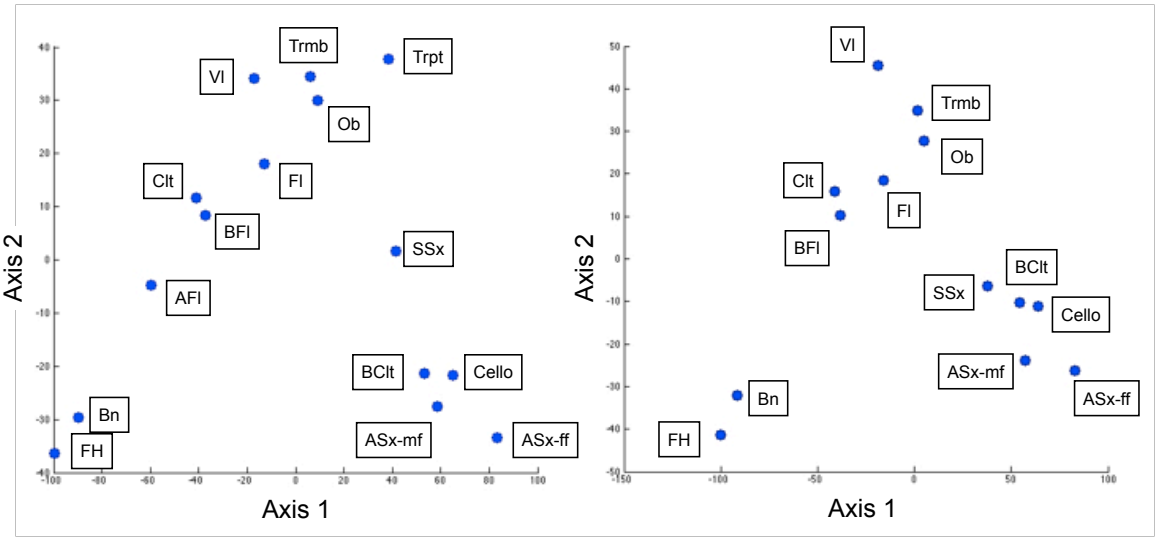


Figure 7.9(a): 15 instruments (left hand column) and 13 instruments (right and column) projected on to axes 1 and 2.

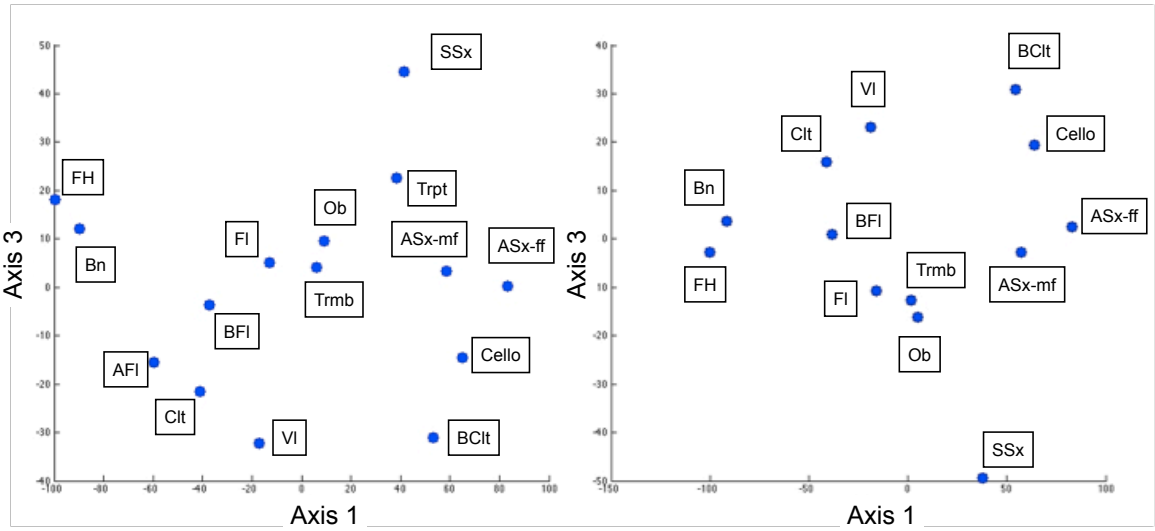


Figure 7.9(b): 15 instruments (left hand column) and 13 instruments (right hand column) projected on to axes 1 and 3.

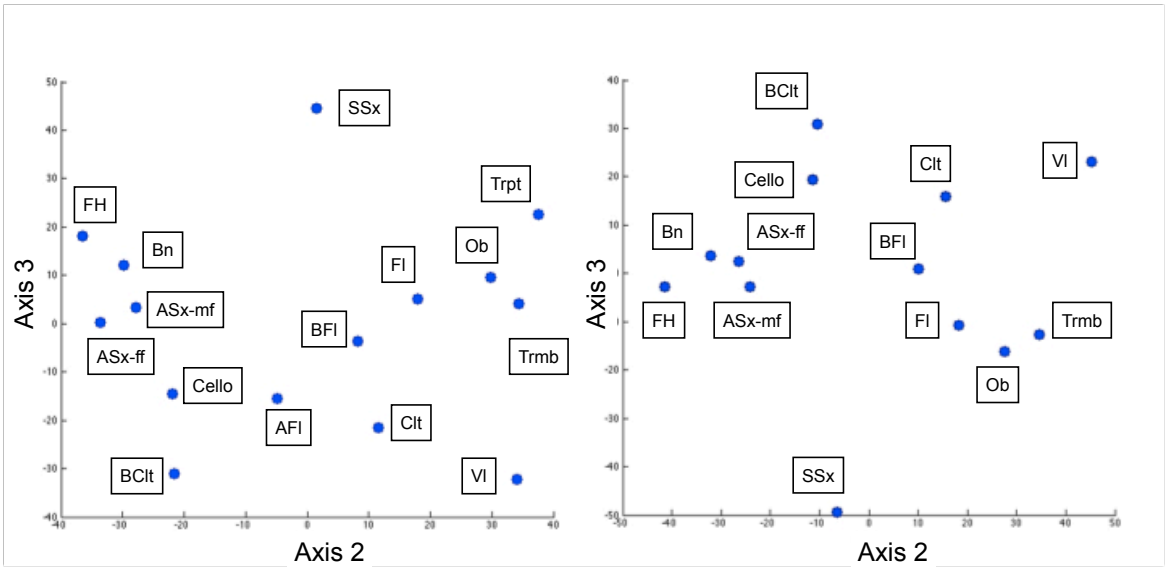


Figure 7.9(c): 15 instruments (left hand column) and 13 instruments (right hand column) projected on to axes 2 and 3.

As can be seen, the removal of two instruments results in a degree of shifting of the relative positions of those remaining, but, broadly speaking, the space remains stable. Note that the third axis of the space generated following the removal of two instruments is reversed.

7.3.5. The seventh dimension – attack time

The six dimensions of the reduced space describe sounds which are dynamically invariant. In order to maintain a degree of methodological continuity with the other two spaces, the seventh axis describes the attack envelope. The attributes are the same as that of the rise time axis of the SCG-EHA space of the previous chapter– i.e. ranging from 0.01 to 0.2 seconds.

7.3.6. Resynthesis of a point in the reduced dimensionality space

Sounds represented in the reduced space **Y** can be auditioned by means of a data recovery process. A given sound can be dynamically generated from a single six-coordinate point in the space as shown in figure 7.10.

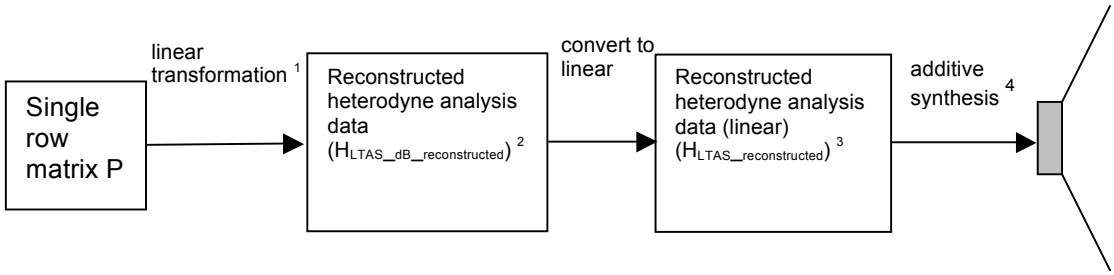


Figure 7.10: Process of reconstructing sounds from reduced space.

1. The single row, 6 column matrix P containing the coordinates of a point in **Y** is transformed using the data obtained from the *procrustes()* function, in order to recover the heterodyne data and to best align it with the original matrix H_{LTAS} .

$$H_{LTAS\_dB\_reconstructed} = (P * R) + T$$

2. The resultant single row, 20 column matrix $H_{LTAS\_dB\_reconstructed}$ contains the long time averaged amplitudes of the harmonics of the desired sound.

3. The elements of this matrix are converted to linear form, as follows :

$$H_{LTAS\_reconstructed} = 10^{\frac{H_{LTAS\_dB\_reconstructed}}{20}}$$

4. This data can be input to an additive synthesis process for playback. The overall envelope (variable rise time and fixed decay envelope is imposed at this point).

7.3.7. Comparison of spectra recovered from the reduced space with original spectra.

The spectra from the resynthesis process were compared with the original spectra; figure 7.11 shows the original and reconstructed heterodyne spectra for the alto flute (spectra for the other instruments are to be found in the appendix). For the most part, the recovered spectra are almost identical to the original spectra – in some cases, the error arising from configuring the data in the reduced space results in negative amplitudes. For the purposes of synthesis, these can be zeroised.

Spectrum from H_{LTAS_dB}	Spectrum from $H_{LTAS\_dB\_reconstructed}$

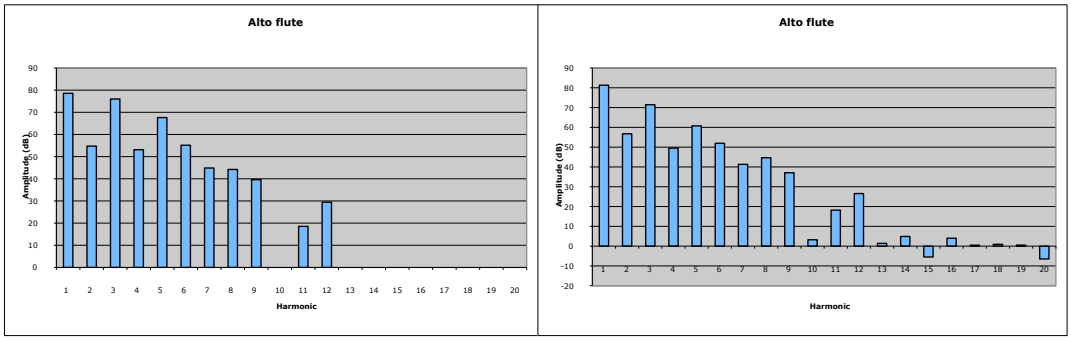


Figure 7.11: Original and reconstructed heterodyne spectra for the alto flute.

7.4. WCL strategies in seven dimensional space

Chapter six described the algorithm behind, and operation of the two strategies, WCL-2 and WCL-7, in a three dimensional attribute space. The strategies are almost unchanged when deployed in a seven dimensional space. The axes are **1..I**, **1..J**,**1..K** , **1..L**, **1..M**, **1..N** and **1..O** ; each represent a single dimension of the MDS solution.

The only important difference here is the addition of four more equations for the seven weighted centroid coordinates $\mathbf{i_C}$, $\mathbf{j_C}$, $\mathbf{k_C}$, $\mathbf{l_C}$, $\mathbf{m_C}$, $\mathbf{n_C}$ and $\mathbf{o_C}$, and which are given below.

$$i_C = \frac{\sum_{x=1}^{N} w_x i_x}{\sum_{x=1}^{N} w_x}, j_C = \frac{\sum_{x=1}^{N} w_x j_x}{\sum_{x=1}^{N} w_x}, k_C = \frac{\sum_{x=1}^{N} w_x k_x}{\sum_{x=1}^{N} w_x}, l_C = \frac{\sum_{x=1}^{N} w_x l_x}{\sum_{x=1}^{N} w_x},$$

$$m_C = \frac{\sum_{x=1}^{N} w_x m_x}{\sum_{x=1}^{N} w_x}, n_C = \frac{\sum_{x=1}^{N} w_x n_x}{\sum_{x=1}^{N} w_x}, o_C = \frac{\sum_{x=1}^{N} w_x o_x}{\sum_{x=1}^{N} w_x}$$

7.5. Procedure

The testing procedure was as described in the previous chapter. Three versions of the software were prepared and were loaded onto three Apple eMac computers. These were as follows:

I: MLS - Multidimensional line search.

II: WCL-2 - Two-alternative forced choice

III: WCL-7 Seven-alternative forced choice

Twenty subjects were used. Each subject was asked to run each test **I** to **III**; again, the order in which the tests were run varied randomly for each subject. The target was the same for all versions.

7.5.1. Multidimensional line search

Each subject was asked to manipulate the seven software sliders shown in figure 7.12, listening to the generated sound until EITHER 'Play sound' had been clicked on sixteen times OR a slider setting was found for which the generated sound was judged to be indistinguishable from the target.

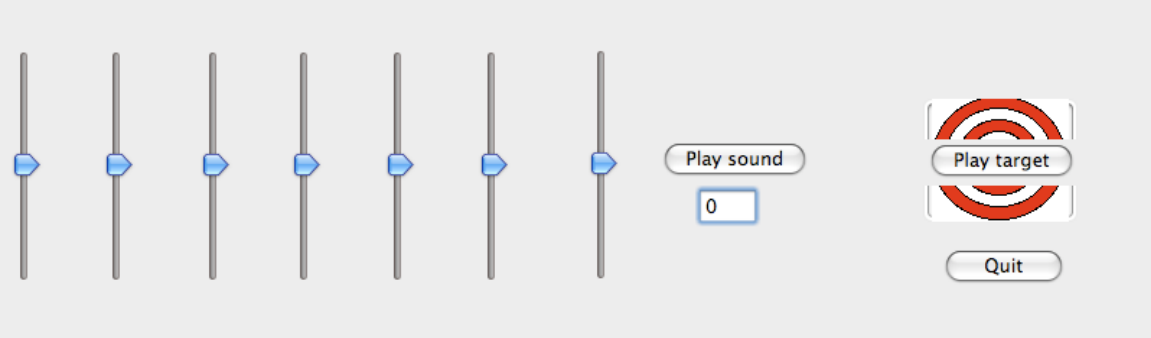


Figure 7.12: Multidimensional line search in a seven dimensional space using seven sliders.

7.5.2. WCL-2 - two-alternative forced choice

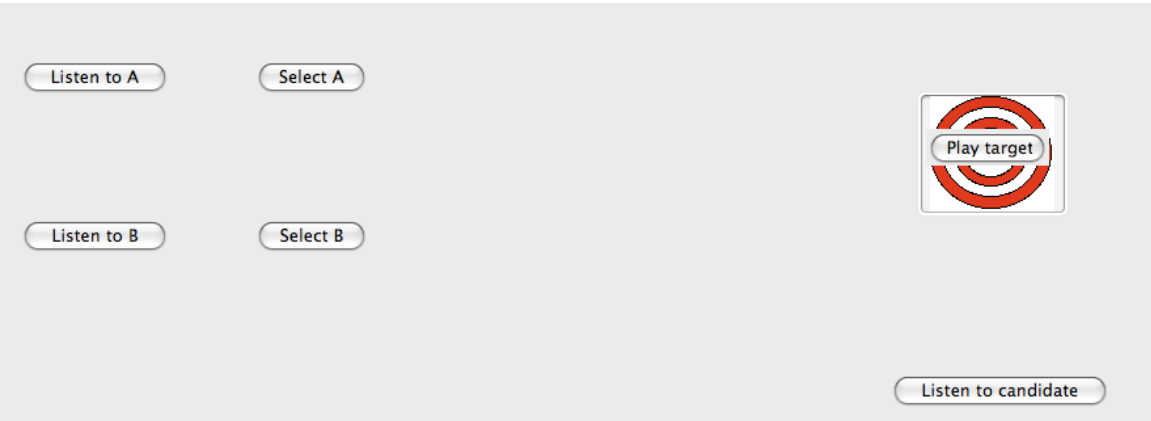


Figure 7.13: WCL-2 – interface for the two-choice algorithm.

The procedure is exactly the same as that used for the three dimensional spaces described in 7.1.1.2 - each subject was asked to listen to the target and then judge which of two sounds A or B more closely resembled it.

7.5.3. WCL-7 - seven-alternative forced choice

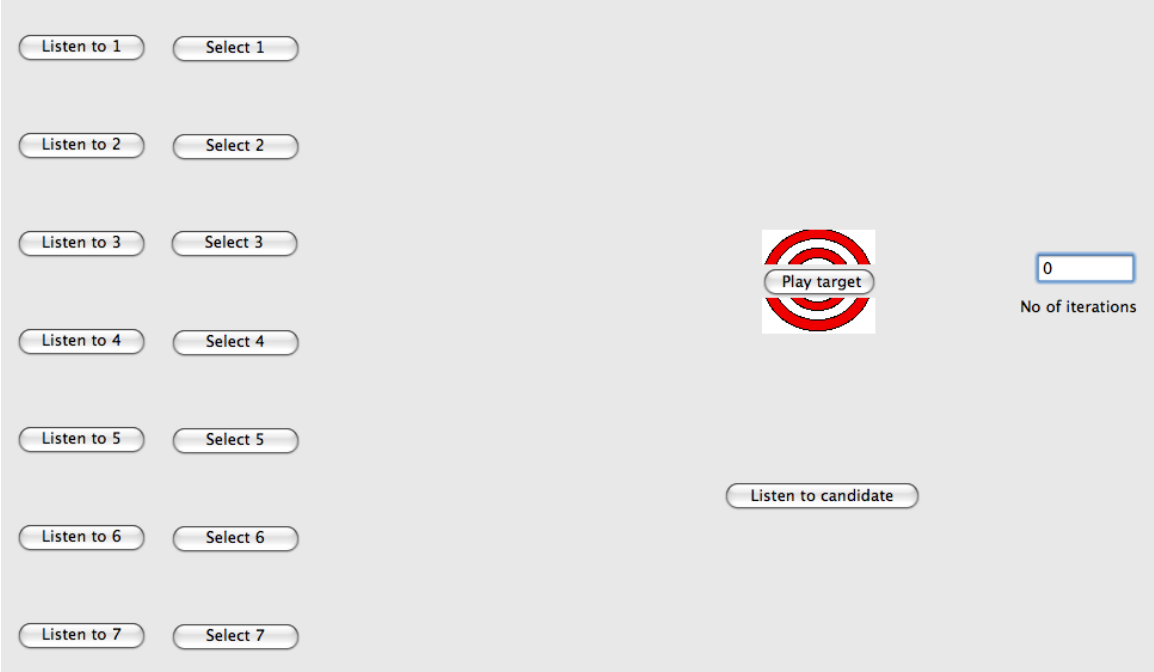


Figure 7.14: WCL-7 – interface for the seven choice algorithm.

The procedure is exactly the same as that used for the three dimensional spaces described in 6.5.1.3. Each subject was asked to listen to the target and then judge which one of the seven sounds heard by clicking on the seven buttons labelled ‘‘Listen to 1”, “Listen to 2”, etc more closely resembled it.

7.6. Results

Again, we start with the results from the multidimensional line search, where subjects are supplied simply with sliders connecting to the axes of the space.

7.6.1. Multidimensional line search

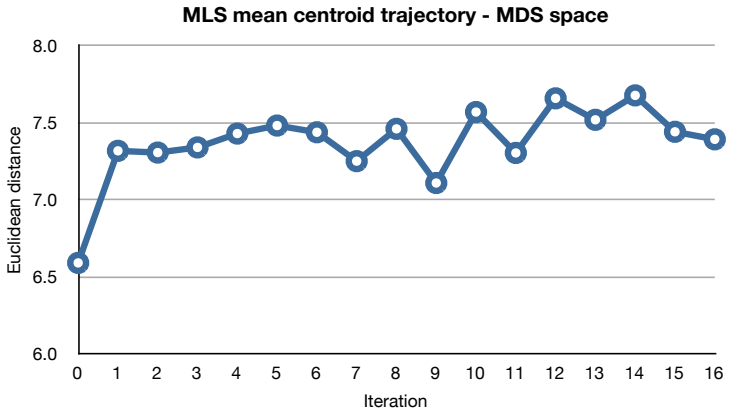


Figure 7.15: Mean weighted centroid trajectory in MDS space using multidimensional line search.

Figure 7.15 shows the averaged trajectory over sixteen iterations for all twenty interactions in the MDS space, and seems to indicate that, overall, the linear search method is not a satisfactory search strategy in this particular attribute space. Inspection of the individual trajectories shows only one example of a subject who was able to use the controls to converge on the target.

The equivalent results from the three dimensional spaces reported in the previous chapter (see section 6.6) included discussion of the mean trajectory as projected on each of the respective axes of the two spaces. This will not be done here, however, for the MDS space. This is because the axes here represent MDS

factors rather than distinct parameters of synthesis or acoustic attributes, and as such are, in themselves, not meaningful.

7.6.2. WCL-2 - two-alternative forced choice

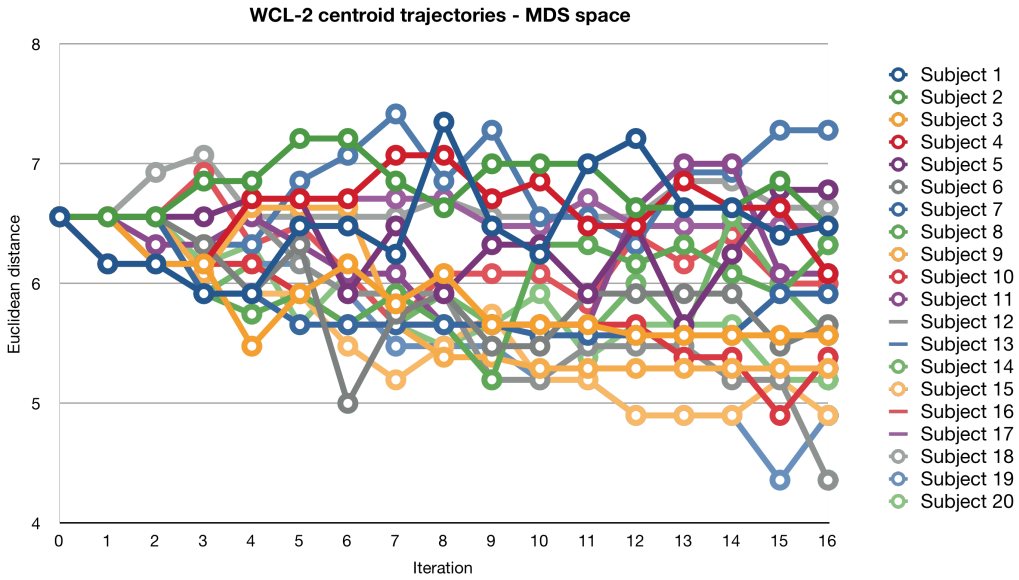


Figure 7.16(a): Mean weighted centroid trajectory in MDS space using WCL-2 strategy.

We turn now to the data from the two-alternative forced choice test in the MDS space. Figure 7.16(a) shows the change in the Euclidean distance in the formant space between the weighted centroid of the probability table and the cell in the probability table corresponding to the target, for all twenty subjects. In comparison with the equivalent trajectories evident in the tests in the previous chapter, there is considerable variation in the trajectories, compared with the equivalent ones for the formant and SCG-EHA spaces discussed in the previous chapter. Figure 7.16(b) shows the mean trajectory followed by the weighted centroid relative to the target.

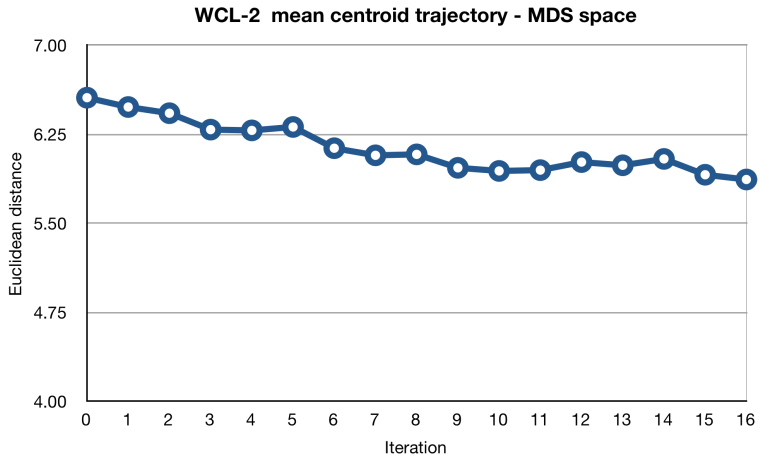


Figure 7.16(b): Mean weighted centroid trajectory in MDS space using WCL-2 strategy.

7.6.3. WCL-7 - seven-alternative forced choice

Finally , we consider the results from the WCL-7 strategy operating within the MDS space . Again, there is considerable variation in the individual trajectories, shown in figure 7.17. Figure 7.18 shows the mean centroid trajectory, which exhibits a steady convergence on the target.

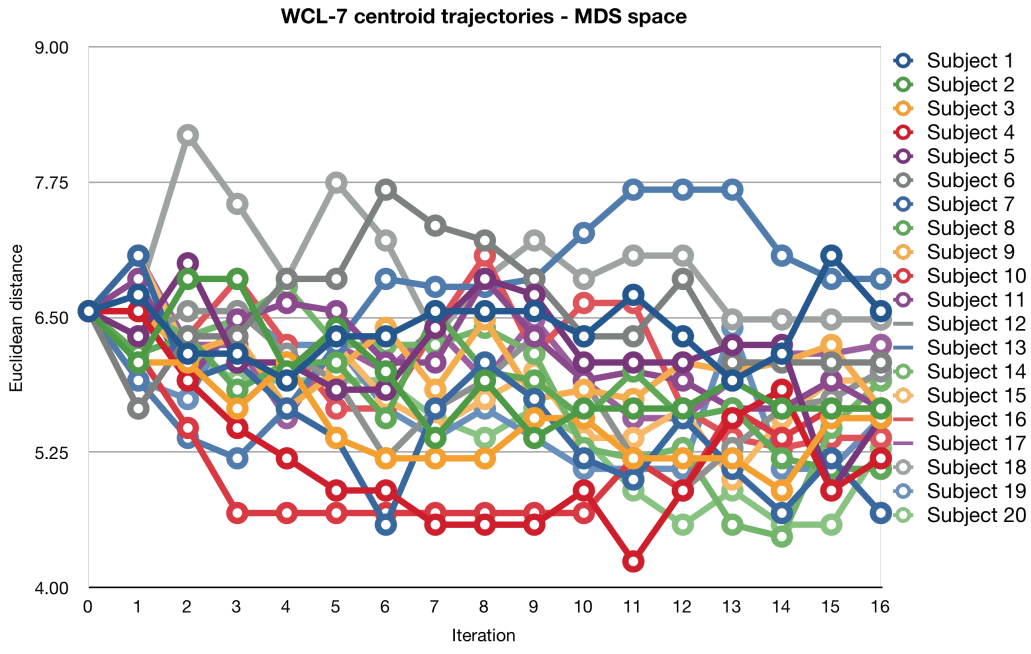


Figure 7.17: Weighted centroid trajectories in MDS space using WCL-7 strategy.

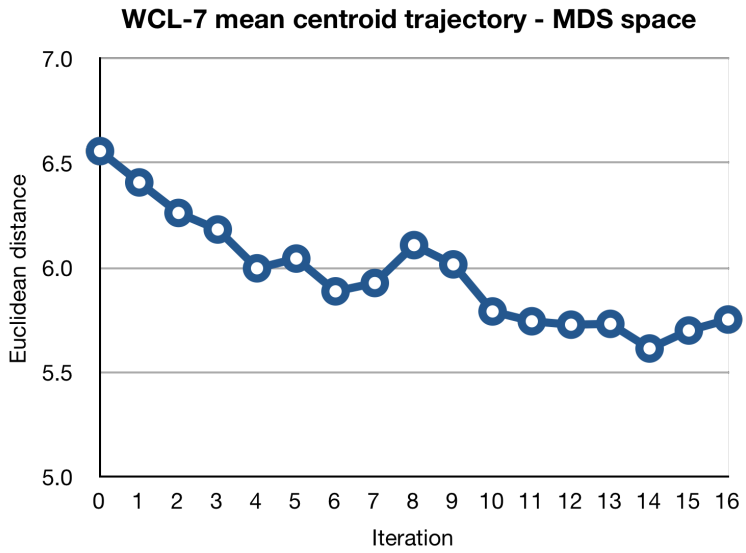


Figure 7.18: Mean weighted centroid trajectory in MDS space using WCL-7 strategy.

7.7. Summary of results

A summary of the results from this chapter, combined with those from the previous chapter is presented in figure 7.19. In order to make possible direct comparison of the results from three attribute spaces that otherwise differed, both in their sizes and in their characteristics, the vertical axis represents the percentage of the Euclidean distance between the target and the initial position of the weighted centroid.

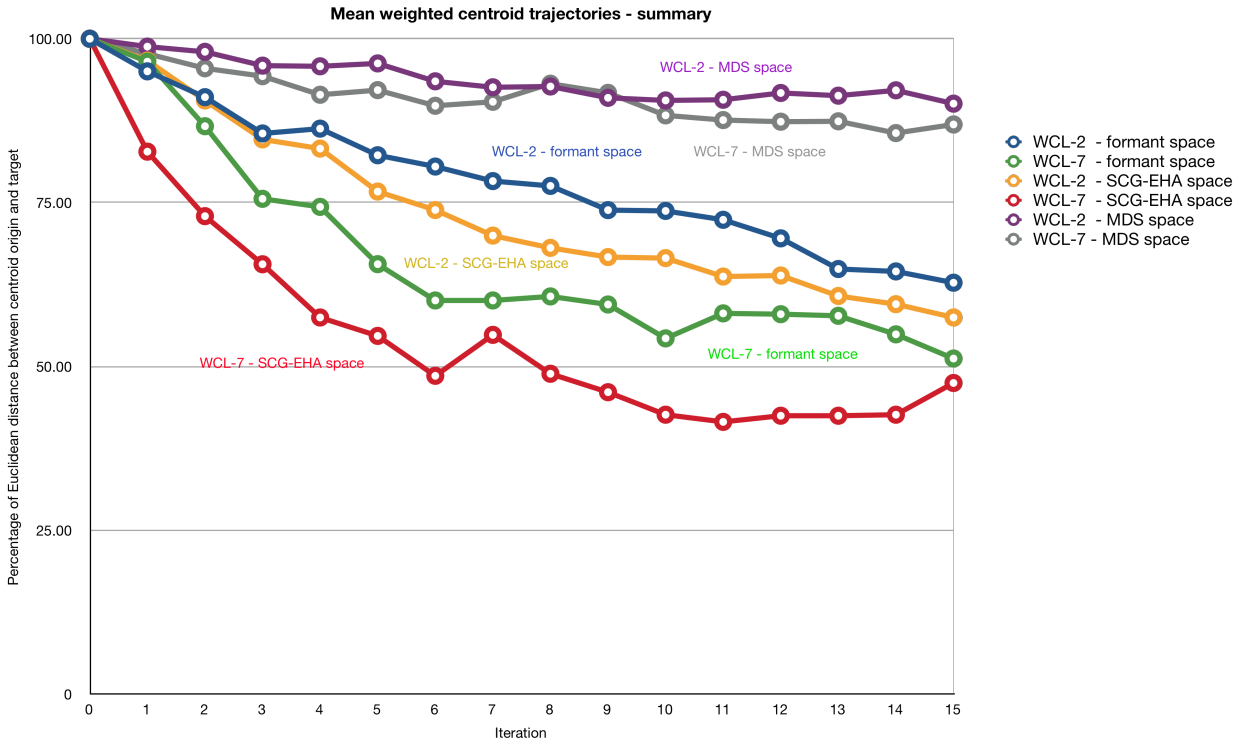


Figure 7.19: Mean trajectories of weighted centroid for WCL-2 and WCL-7 strategies in three different attribute spaces.

While it should be borne in mind that, in all cases, there was considerable variation in individual subject performance, the six mean weighted centroid trajectories from the WCL-2 and WCL-7 search strategies in the three spaces all show, to a greater or lesser extent, a convergence on the target. Two observations can be made from the above results.

Firstly, the gradients of the two traces representing the weighted centroid mean trajectory in the seven-dimensional MDS space are considerably shallower than those in either of the two three-dimensional spaces. One probable reason for this is the greater difficulty of the task; a seven dimensional space is clearly more difficult to navigate than a three dimensional one. Another possible reason is the far greater inertia of the MDS space probability table (consisting of $7^7 = 823543$ cells) relative to that of the formant and SCG-EHA probability tables (1690 and 1815 cells, respectively), which would cause a slower shift of the weighted

centroid across the space. This might be addressed by increasing the factor by which probability values are updated on each iteration corresponding to a sound which is 'closer' to a chosen sound.

Secondly, in each of the three attribute spaces, the WCL -7 strategy, in which subjects were asked to choose from seven probes, produced a swifter convergence (expressed as the number of subject iterations) on the target than the WCL-2 strategy, where only two probes were offered. This was observable in a number of individual subject performances, as well as in the overall graph, and is an interesting result. The task of critically evaluating seven, rather than two probes imposes on the subject a greater cognitive load and it had been speculated that this would result in a slower (or even zero) rate of convergence. Again, the gradient is likely to be highly sensitive to the value of the multiplication factor.

It should be emphasised, however, that the metric used here is the number of iterations, not the elapsed time or the number of individual actions (i.e. mouse clicks) required to audition the seven probes. Several subjects reported that the WCL-7 task was more difficult than the WCL-2 task; and although this was not measured, it was noticeable that the time required by a number of subjects to complete the task was significantly greater in the case of the WCL-7 task than for either of the other two. Similarly, the minimum number of mouse clicks required to listen to seven probes is obviously greater than that required for two.

The final chapter will consider these points in greater detail, together with other possible modifications and techniques which could be applied to improve the convergence gradient in a multidimensional space.

Chapter 8 - Conclusion

8.1. Introduction

This thesis has

- examined problems and issues relevant to the design of an intuitive user interface for musical timbre, highlighting the gulf between user and system languages;
- reviewed the body of research into timbre, both as musical resource and as psychoacoustical phenomenon;
- discussed current approaches for bridging the user-system language gap;
- presented an interface, based on weighted centroid localization, for searching suitably configured timbre spaces.

This chapter summarises the contributions made to the disciplines which it draws on, and goes on to consider limitations, the kind of synthesis engine which could be driven by the WCL search strategy, possible practical implementations and areas for further work.

8.2. Contributions of this thesis

The thesis is located within three different disciplines – Music Computing, HCI and psychoacoustics, and has made the following contributions to them.

8.2.1. To Music Computing

The main work of this thesis, outlined in section 1.2 and explored at length in chapters six and seven, is a contribution to the field of music computing in that

it proposes a means of searching timbre space based on the iterated updating of a probability table. While having much in common with current work using genetic algorithms, the WCL search strategy is a distinctly different means of searching a timbre space.

Whereas other approaches to the problem (such as genetic algorithms) have employed search methods based on some form of parameter optimisation, the approach taken here has been to redefine the problem as one of localisation. It is claimed that a weighted centroid localisation method, where the weights are probability values which are iteratively updated, offers a useful method of searching a particular subset of synthesis spaces whose axes are perceptually linear, and in which relative Euclidean distances between sounds inhabiting the space broadly correspond to perceptual judgments of similarity/dissimilarity. In such spaces where the fitness contour is simple, the GA processes of crossover and mutation, in which new and very different candidate sounds are generated, are likely to disrupt, rather than support the search.

The empirical work suggests that the WCL method performs significantly better in relatively simple three dimensional spaces (in this case the formant space and the SCG-EHA spaces) than in spaces where the dimensionality is greater (the MDS space) – a result which is, in itself, not surprising; but in all three spaces, the two versions of the WCL performed better than the MLS strategy (where subjects had direct access to the dimensions of the space).

The subsidiary proposal, whose purpose was outlined in section 1.2 and discussed in chapter four, adds to the existing literature on music computing by proposing a set of criteria for an ideal *n*-dimensional attribute space which

functions usefully as a vehicle for search strategies such as those described in this thesis.

8.2.2. To HCI

As noted in the introduction to chapter one, the synthesizer user interface (as distinct from real-time performance interfaces) has received relatively little attention from the HCI research community. The study of the synthesizer user interface, whose objectives are given in section 1.2 and which is presented in chapter two, contributes to the HCI literature by:

- proposing a taxonomy of interaction styles for manipulating timbre in synthesizers which has three categories - *parameter selection in a fixed architecture*, *architecture specification and configuration*, and *direct specification*;

- identifying the interaction styles which are most suited to different synthesis methods. It was noted that those synthesis methods whose parameters map more readily to measurable acoustic attributes seem to be best implemented by *direct specification* methods, whereas more abstract synthesis techniques require *fixed architecture* or *architecture specification* implementations.

- analyzing a number of hardware and software synthesis implementations representative of these three categories, concluding that, in terms of the number of user actions required to complete a sound specification task, *direct specification* interfaces are more usable and intuitive than *fixed architecture* or *architecture specification* interfaces.

8.2.3. To psychoacoustics

The work on subject perception of Euclidean distances in timbre space, the aim of which was outlined in section 1.2, and which was presented in chapter five, builds on previous research by Ehresman and Wessel (1978), McAdams and Cunible (1992) and Toiviainen, Kaipainen and Louhivuori (1995) in examining the extent to which Euclidean distances between sounds disposed in a timbre space are reflected in perceptual distances. While it cannot be claimed that such a link will exist in all possible timbre spaces, for those spaces where it can be demonstrated (and for this reason are suitable vehicles for the WCL search strategy), the results show clear evidence of a positive correlation between the perceptual granularity of different parts of the space and subjects' ability to perceive relative Euclidean distances in those varying regions.

8.3. Limitations of the research

Before considering how the WCL strategy might be implemented, we consider here a number of limitations, some of which have been touched on in previous chapters, but which are summarised here, together with proposals for addressing them.

First of all, the sounds inhabiting all three spaces are spectrally and dynamically invariant (although the SCG-EHA and MDS spaces include a variable attack time dimension). Clearly, for the strategy to be a useful tool for timbral shaping, this would need to be addressed; however, to implement this in the interface presented here would require a more complex mapping of the search space to the probability space; this is a line of research to be pursued in the future.

As was stated in the introduction to chapter five, the WCL search strategy (in either of its forms) is of very limited use in synthesis environments where the mapping between the perceptual space and the synthesis space is not straightforward (in the case of FM synthesis, for example). This need not necessarily preclude other synthesis methods; but, as discussed below in the section on appropriate synthesis methods, the processing would be computationally more expensive.

The discussion at the end of the previous chapter noted that the metric used was simply the number of iterations; if elapsed time and/or the number of user actions (mouse clicks) is used as a performance indicator, it is by no means clear that the WCL-7 strategy is the best of the three in any of the spaces it was deployed. The reason is that the interaction process is currently slow because of the long response time for each iteration, caused by the updating of the probability table. Because the number of cells to be updated increases exponentially with the number of search space dimensions, it is particularly noticeable in the seven dimensional MDS space discussed in the previous chapter. Slow response times for each iteration may well be addressed by rewriting the software in (for example) C++ or Objective-C, with the part of the code which implemented the update loop being rewritten in Assembler.

However, the number of iterations required to achieve a significant degree of convergence with the target is also high. Essentially, this is the 'bottleneck' problem, characteristic of interactive GAs and discussed in chapter four. Methods of addressing both this issue and that of long response times are considered in the section on practical implementation (section 8.5).

All three spaces investigated in this thesis have been constructed such that distances between the sounds in the space are Euclidean distances, rather than being based on any other metric. This is justifiable, as we are primarily interested in relative distances, both real and perceived, rather than in the perceptual properties of the spaces themselves. However, there are other metrics that can be used, which were explored by McDermott, Griffith *et al* (2005) and reviewed in chapter four of this thesis. In particular, neither the WCL process itself, nor the spaces in which it was tested took account of the non-linearity of human hearing. That the sensitivity of the hearing mechanism varies with frequency is well known (Fletcher and Munson, 1933); this is built into audio compression algorithms such as MPEG-1 Audio Layer 3 (MP3). However, neither the search spaces navigated by the search strategies, nor the strategies themselves have incorporated a perceptual model which reflected this. It would be of interest to establish whether the WCL search strategy performed significantly better in such spaces.

As discussed in chapter six (section 6.3.2.5.1), the multiplication factor of √2 used to update the cells of the probability table in the WCL-2 strategy is not necessarily optimal. This is also true of the factor used in the WCL-7 strategy, whose value was inversely related to the distance from the chosen probe. In the WCL-7 strategy, the gradient of cell values in the probability table is linear (related to distance from chosen probe); it would be of interest to ascertain whether better results might be obtained if it was (for example) exponential.

However, in both cases, increasing the value carries with it a penalty; the probability table would not so quickly 'recover' from 'incorrect' user judgments. Conversely, decreasing it provides insufficient 'reward' for 'correct' ones, and would effectively prolong the interaction. One way of arriving at the optimum value might be through the use of a search algorithm. Repeated automated runs of

the algorithm with simulated user input (such as those done for the 'control' tests in chapter six) could be made with different multiplication values, with the optimum value being the one that tended to result in the lowest number of iterations – that is to say, the steepest gradient in the Euclidean distance between the weighted centroid and the target.

Finally, the caveat noted in chapters two and six is reiterated here. In order to test the strategies, a target sound was provided for the subjects, whereas the ultimate purpose, of course, is to provide a user interaction which converges on a target which is imaginary. The assumption is, firstly, that the imagined sound actually exists in the space and can be reached; and secondly, that it is stable – the user's imagined sound does not change.

8.4. Further work

We conclude by discussing possible practical implementations of the WCL strategy, the type of synthesis engine that would be best suited to such implementations and other directions for future work.

8.4.1. Synthesis engines appropriate to the WCL strategy

The first of the three attribute spaces explored in chapters six and seven (whose axes were formant centre frequencies) were generated for the purposes of the study by simple formant synthesis; those of the second (spectral centroid, even harmonic attenuation and attack time) and of the third (attack time and the six multidimensional scaling factors) by additive synthesis. Both these methods fall into the *spectral model* category of synthesis, and are suitable for the WCL search algorithm because their parameters can be used to form search spaces in which the

mapping between relative perceptual distances and Euclidean distances is relatively straightforward. One possible way of using an additive synthesis engine with the WCL algorithm might be as part of an analysis-resynthesis system, in which a set of samples was analysed using heterodyne analysis techniques, and a search space of low dimensionality generated from this data using MDS (as shown in chapter seven).

The extent to which other synthesis engines could also be used to generate search spaces for the WCL strategy is, as already stated, dependent on whether a mapping of relative perceptual distances and Euclidean distances can be demonstrated. A space derived from a simple *source-filter* or *subtractive synthesis* structure (also an example of the spectral model category) is potentially usable as a search space for the WCL algorithm; changes in single parameter values, such as the cut-off frequency, for example, are both clearly audible and produce a proportional degree of timbral change. It is not obvious, however, that this would be the case in a more intricate structure, consisting of (for example) two or three voltage controlled oscillators whose outputs were filtered in different and complex ways. The one-to-one mapping of a parameter from this structure to a single timbral attribute is not likely to be straightforward; neither would distances in the synthesis space necessarily correspond to those in the perceptual space. If, on the other hand, such a mapping could be demonstrated for a complex subtractive synthesis space, MDS techniques could be employed to reduce the dimensionality. (It should be noted that such a space could only be implemented in a *fixed architecture* interface (as discussed in section 2.6.3.1.1), where the dimensionality is fixed, rather than one where individual components can be added or removed from the structure.)

As noted in chapter two (section 2.4.3), the source-filter synthesis method has features in common with *physical modelling* synthesis, in that sound is viewed as the output of a network of functional components. Thus, the argument made in the previous paragraph applies here as well; the WCL strategy could only be usefully deployed if distances in the physical modelling synthesis space more or less corresponded to distances in the perceptual space. This is not to say, however, that WCL could not be applied at all to these synthesis engines; but such a system would be computationally far more complex and require some form of interpretative layer. As noted in the introduction to chapter five, search strategies such as genetic algorithms might be better suited.

Synthesis engines which come under the *abstract model* and *processed recording* categories (FM and granular synthesis, for example) are even less well suited to the WCL strategy. Few of their parameters map easily and linearly to perceptual parameters. Again, the search spaces of methods of synthesis such as these might be more effectively navigated by genetic algorithms.

8.4.2. Practical implementation

The advantage of the WCL search strategy stated in chapter six (section 6.3.2.1) was that the user did not need to be familiar with either the parameters of the synthesis engine or the acoustical attributes of the sound in order to effect change in the sound. In practice, an interface in which the user simply makes choices from two (or seven) candidates is unnecessarily restrictive. It is generally good practice in user interface design to allow for different levels of expertise, to provide shortcuts and, in general, to build flexibility into the interface (Nielsen, 1994). This applies no less in the sound synthesis domain; Polfreman and Sapsford-Francis's study, noted in chapter two, recommended the provision of

more than one level of interaction and the hiding of unwanted levels of complexity in the design of computer music systems (Polfreman and Sapsford-Francis, 1995).

This being the case, there is potential for an implementation of the WCL strategy as a sub-system within a system offering a wider range of synthesis functions. A template for this is the *Patch Mutator* (discussed in chapter four). This IGA-based method of searching the synthesis space is integrated into the software patch editor for the Nord Modular G2 hardware synthesizer.

Other possibilities include the incorporation of the strategy as a VST or AU plug-in for a synthesis package such as *AudioMulch*. With the rapid emergence of touch screen technology, a user interface could be designed which embodied the WCL strategy, and which communicated via Open Sound Control (OSC) or Bluetooth with a hardware or software synthesizer. Any of these implementations would offer the user choice and flexibility in the synthesis tools available at any given time.

As noted in the previous section, the WCL search strategy presented in this thesis only allows the user to make choices of sounds which are spectrally and dynamically invariant; however, sound objects of arbitrary complexity with respect to time could nevertheless be generated by using the WCL method to create breakpoints, or 'snapshots' of a dynamically evolving sound. The sound in its entirety could then be constructed by interpolation between these breakpoints.

However the strategy is implemented, the user-system interaction needs to be accelerated if the strategy is to be practically usable; the system is both too limited in scope and too slow in operation for 'real world' use. Possible directions

for new work which could be used in a practical implementation and which would address these problems will now be considered.

Convergence on the target might be significantly accelerated if the user, instead of being offered two or more probes for consideration, is provided with a slider which offers sounds which are graduated interpolations between two points in the space, or alternatively a two dimensional slider which interpolate between four points. Very much the same technique is proposed and described in McDermott *et al* (2007) as a means of selection; what is proposed here is an adapted version of it. On the face of it, this may seem like a reversion to the multidimensional line search (MLS) slider interface used in chapters six and seven. However, there is an important difference. Each of the sliders used in the MLS interface was tied to one axis of the search space – a formant centre frequency (in the formant space), attack time, spectral centroid and even harmonic attenuation (in the SCG-EHA space), or one of the axes generated in the MDS space. What is proposed here is a slider which is not tethered to any one of these axes, but is dynamically attached to a vector which joins two probes whose positions in the space are updated on each iteration. A two dimensional slider could be similarly used for a vector which joined three probes.

It is the nature of interactive search that some searches prove not to be fruitful, or the target changes during the course of the interaction (because the user has changed his/her mind). The discussion in chapter six of the 'stability' of the imagined target sound is relevant here. This being the case, the interface could easily incorporate a 'backtrack' feature, by use of which previous iterations could be revisited, and new choices made. One useful method of backtracking was afforded by the *Mutasynth* system; it featured a method by which stored sounds, visually represented by a mnemonic whose shape depicted the synthesis parameter values, could be recalled. The WCL strategy could similarly recall

promising candidate sounds which had been previously stored, allowing comparison with the current generation of probe sounds.

Another direction which could prove fruitful is to provide the user with the means of rating two or more probes for perceived similarity to the target, (rather than simply selecting one). The probability space could then be given an additional weighting based on the relative rating of the probes, which in turn might result in a swifter convergence of the weighted centroid on the target.

The discussion, in chapter six (section 6.3.2.1) of the WCL interface emphasised that it afforded engagement with the sound itself, rather than with a visual representation of it. Earlier trial versions of the software (not discussed in this thesis) did, in fact, include a visual representation of the probability space, in which the cell values were colour-coded. The probability space was thus represented as a contour map, from which users were able to see areas of higher probability in the search space. Interestingly, during pilot testing, subjects found the visual element to be a distraction; they stopped listening critically to the sounds and instead, selected those probes which appeared to be closer to these 'high probability' areas. In this way, a dialogue was established in which the software was driving user choices, rather than the other way round.

However, the idea of including some form of visual representation could, nevertheless, be revisited. If, for example, the trajectory of the weighted centroid through the *n*-dimensional space is more or less a straight line, this could be graphically represented, enabling the user to explore sounds which exist on a extrapolation of this trajectory.

8.4.3. Other directions for future research

In the course of this thesis, the WCL strategy has been discussed in the context of other search algorithms for synthesis – in particular, those which are based on knowledge based systems (KBS), which exploit genetic programming techniques, or which make use of interactive and non-interactive genetic algorithms. However, no attempt has been made here to compare the effectiveness (expressed in elapsed time, number of user actions or iterations) of any of these techniques with that of the WCL strategy; nor, in the course of this research, has any similar comparative evaluation of these techniques come to light (at least in the domain of sound synthesis). Such a study would be a useful guide to further work in this area.

More generally, the current literature on usability evaluation of synthesizers, represented by the work of Jaffe, Tolonen *et al* and others (reviewed in chapter two of this thesis) is limited, and there is scope for further research in this area. In particular, there is a need for a rigorous methodology for usability evaluation of audio hardware and software and for more work in this area in general.

References

Abramson, B. (1994) The design of belief network-based systems for price forecasting. *Computers and Electrical Engineering*, 20(2), 163-180.

American National Standards Institute (1973). *USA standard psychoacoustical terminology*. New York: American National Standards Institute.

Andreassen, S., Woldbye, M., Falck, B. & Andersen, S. K. (1987) MUNIN: A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, Milan, Italy, 366-372.

Arfib, D. (1979) Digital synthesis of complex spectra by means of multiplication of non-linear distorted sound waves. *Journal of the Audio Engineering Society*, 27(10), 757-768.

Ashley, R. (1986) A Knowledge-Based Approach to Assistance in Timbral Design. *Proceedings of the 1986 International Computer music Conference*, The Hague, Netherlands, 11-16.

Aures, W. (1985) Ein Berechnungsverfahren der Rauhigkeit. *Acustica*, 58, 268-280.

Bäck, T. (1996) *Evolutionary Algorithms in Theory and Practice*. New York: Oxford University Press.

Balzano, G. J. (1986) What are musical pitch and timbre? *Music Perception*, 3(3), 297-314.

Barker, A. (2000) *Scientific Method in Ptolemy's 'Harmonics'*. Cambridge and New York: Cambridge University Press.

Beauchamp, J. (1969) A computer system for time-variant harmonic analysis and synthesis of musical tones. In: H. von Foerster & J. W. Beauchamp, ed. *Music by Computers*. New York: Wiley.

Beauchamp, J. (1975) Analysis and synthesis of cornet tones using non-linear interharmonic relationships. *Journal of the Audio Engineering Society*, 23(10), 718-795.

Beauchamp, J. (1981) Data reduction and resynthesis of connected solo passages using frequency, amplitude, and `brightness' detection and the nonlinear synthesis technique. *Proceedings of the 1981 International Computer Music Conference*, Denton, Texas, 316-323.

Bencina, R. & Burk, B. (2004) PortAudio: An API for Portable Real-Time Audio. In: K. Greenebaum & R. Barzel, ed. *Audio Anecdotes*. A K Peters.

Berger, K. W. (1964) Some factors in the recognition of timbre. *Journal of the Acoustical Society of America*, 36(10), 1888-1891.

Beyer, H-G. & Schwefel, H-P. (2002) Evolution Strategies: A Comprehensive Introduction. *Natural Computing*, 1(1), 3-52.

Biles, J. A. (1994) GenJam: A Genetic Algorithm for Generating Jazz Solos. *Proceedings of the 1994 International Computer Music Conference (ICMC '94)* 131-137.

Bismarck, G. von (1974) Sharpness as an attribute of the timbre of steady sounds. *Acustica*, 30(3), 159-172.

Bismarck, G. von (1974) Timbre of steady sounds: a factorial investigation of its verbal attributes. *Acustica*, 30(3), 146-158.

Blumenthal, J., Grossmann, R., Golatowski, F. & Timmermann, D. (2007) Weighted Centroid Localization in Zigbee-based Sensor Networks. *WISP 2007. IEEE International Symposium on Intelligent Signal Processing* 1-6.

Breese, J., Horvitz, E., Peot, M., Gay, R. & Quentin, G. (1992) Automated decision-analytic diagnosis of thermal performance in gas turbines. *Proceedings of the International Gas Turbine and Aeroengine Congress and Exposition*, Cologne, Germany.

Butler, D. (1992) *The musician's guide to perception and cognition*. New York: Schirmer Books.

Bylstra, M. & Katayose, H. (2005) Painting as an Interface for Timbre Design. In: ed. *Entertainment Computing - ICEC 2005*. Berlin/Heidelberg: Springer**,** 303-314.

Caclin, A., McAdams, S., Smith, B. K. & Winsberg, S. (2005) Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, 118(1), 471-482.

Caetano, M., Manzolli, J. & von Zuben, F. J. (2005) Interactive Control of Evolution Applied to Sound Synthesis. *Proceedings of the 18th International Florida Artificial Intelligence Research Society (FLAIRS)*, Clearwater, EUA.

Chadabe, J. (1997) *Electric Sound*. New Jersey: Prentice Hall.

Cheung, N. M. & Horner, A. (1996) Group synthesis with genetic algorithms. *Journal of the Audio Engineering Society*, 44(3), 130-147.

Chion, M. (1983) *Guide des objets sonores*. Paris: INA-GRM/Buchet-Chastel.

Choudhury, T., Rehg, J. M., Pavlovic, V. & Pentland, A. (2002) Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection. *Proceedings of the 16th International Conference on Pattern Recognition* 789-794.

Chowning, J. (1973) The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society*, 21(7), 526-534.

Clark, M., Robertson, P. & Luce, D. (1964) A preliminary experiment on the perceptual basis for musical instrument families. *Journal of the Audio Engineering Society*, 12(3), 199-203.

Computer Music (2004) The CM Guide to FM Synthesis. *Computer Music,* Sept. 2004.

Cox, T. F. & Cox, M. A. A. (2001) *Multidimensional scaling*. London: Chapman & Hall/CRC.

Dahlstedt, P. (2001) Creating and Exploring Huge Parameter Spaces: Interactive Evolution as a Tool for Sound Generation *Proceedings of the 2001 International Computer Music Conference*, Havana, Cuba, 235-242.

Dahlstedt, P. (2007) Evolution in creative sound design In: E. R. Miranda & J. A. Biles, ed. *Evolutionary Computer Music*. London: Springer-Verlag**,** 79-99.

Dahlstedt, P. (2009) Thoughts on Creative Evolution: A Meta-generative Approach to Composition. *Contemporary Music Review*, 28(1), 43-55.

Darke, G. (2005) Assessment of timbre using verbal attributes. *Conference on Interdisciplinary Musicology (CIM05)*, Montreal, Canada,

Dawkins, R. (1986) *The Blind Watchmaker: why the evidence of evolution reveals a universe without design*. New York: Norton.

Dawkins, R. (1988) The Evolution of Evolvability. In: *Artificial Life, SFI studies in the Sciences of Complexity*. Addison Wesley.

Disley, A. & Howard, D. (2003) Timbral semantics and the pipe organ. *Proceedings of the Stockholm Music Acoustic Conference* 607-610.

Disley, A., Howard, D. & Hunt, A. (2006) Timbral description of musical instruments. *9th International Conference on Music Perception and Cognition*, Bologna, 61-68.

Dissard, P. & Darwin, C. J. (2001) Formant frequency matching between sounds with different bandwidths and on different fundamental frequencies. *Journal of the Acoustical Society of America*, 110(1), 409-415.

Dix, A., Finlay, J., Abowd, G. & Beale, R. (1998) *Human-Computer Interaction*. London: Prentice Hall.

Donnadieu, S. (2007) Mental Representation of the Timbre of Complex Sounds. In: J. W. Beauchamp, ed. *Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music*. Springer**,** 272-319.

Ehresman, D. & Wessel, D. L. (1978). *Perception of Timbral Analogies*. Technical report 13, IRCAM.

Erickson, R. (1975) *Sound Structure in Music*. Berkeley and Los Angeles: University of California Press.

Ethington, R. & Punch, B. (1994) SeaWave: A System for Musical Timbre Description. *Computer Music Journal*, 18(1), 30-39.

Faure, A., McAdams, S. & Nosulenko, V. (1996) Verbal correlates of perceptual dimensions of timbre. *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC4)*, McGill University, Montreal, Canada.

Fernandes, G. & Holmes, C. (2002) Applying HCI to Music-Related Hardware. *Conference on human factors in computing systems (CHI 2002)* 870-871.

Flanagan, J. L. (1955) A Difference Limen for Vowel Formant Frequency. *Journal of the Acoustical Society of America*, 27(3), 613-617.

Flanagan, J. L. (1957) Estimates of the maximum precision necessary in quantizing certain dimensions of vowel sounds. *Journal of the Acoustical Society of America*, 29(4), 533-534.

Flanagan, J. L. (1972) *Speech analysis synthesis and perception*. Berlin: Springer Verlag.

Flanagan, J. L. & Saslow, M. G. (1958) Pitch Discrimination for Synthetic Vowels. *Journal of the Acoustical Society of America*, 30(5), 435-442.

Fletcher, H. & Munson, W. A. (1933) Loudness, its definition, measurement, and calculation. *Journal of the Acoustical Society of America*, 5(2), 82-108.

Freedman, M. D. (1965) *A technique for analysis of musical instrument tones*. Ph.D thesis. University of Illinois.

Freedman, M. D. (1967) Analysis of musical instrument tones. *Journal of the Acoustical Society of America*, 41(4A), 793-806.

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675-701.

Fritts, L. (1997). *The University of Iowa Electronic Music Studios - Musical Instrument Samples*. Retrieved 9th June, 2008, from http://theremin.music.uiowa.edu/MIS.html.

Gabor, D. (1947) Acoustical quanta and the theory of hearing. *Nature*, 159, 591-594.

Gagne, J. P. & Zurek, P. M. (1988) Resonance frequency discrimination. *Journal of the Acoustical Society of America*, 83(6), 2293-2299.

Garcia, R. A. (2001) Growing sound synthesizers using evolutionary methods. In: E. Bilotta, E. R. Miranda, P. Pantano & P. Todd, ed. *Proceedings of ALMMA 2002: Workshop on artificial life models for musical applications*. Cosenza, Italy: Editoriale Bios, 99-107.

Gaver, W. W. (1993) How do we hear in the world? Explorations in ecological acoustics. *Ecological Psychology*, 5(4), 285–313.

George, W. H. (1954) A sound reversal technique applied to the study of tone quality. *Acustica*, 4, 224-225.

Giannakis, K. (2001) *Sound Mosaics : A graphical user interface for sound synthesis based on auditory-visual associations*. Ph.D thesis. Middlesex University.

Giannakis, K. (2006) A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound*, 11(3), 297–307.

Gibson, J. J. (1966) *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979) *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin.

Goldberg, D. E. (1989) *Genetic algorithms in Search, Optimization and Machine Learning* Redwood City, CA: Addison-Wesley.

Gounaropoulos, A. & Johnson, C. (2006) Synthesising Timbres and Timbre-Changes from Adjectives/Adverbs. In: *Applications of Evolutionary Computing*. Berlin / Heidelberg: Springer**,** 664-675.

Grey, J. M. (1975) *An exploration of musical timbre*. Ph.D. Stanford University.

Grey, J. M. (1977) Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61(5), 1270-1277.

Grey, J. M. & Gordon, J. W. (1978) Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63(5), 1493-1500.

Grey, J. M. & Moorer, J. A. (1977) Perceptual evaluations of synthesized musical instrument tones. *Journal of the Acoustical Society of America*, 62(2), 454-462.

Gu, Y., Peiris, D., Crawford, J., McNicol, J., Marshall, B. & Jefferies, R. (1994) An application of belief networks to future crop production. *Proceedings of the 10th Conference on Artificial Intelligence for Applications*, San Antonio, Texas, 305-309.

Hajda, J. M., Kendall, R. A., Carterette, E. C. & Harshberger, M. L. (1997) Methodological issues in timbre research. In: I. Deliège & J. Sloboda, ed. *The Perception and Cognition of Music*. London: Psychology Press**,** 253-306.

Handel, S. (1989) *Listening*. Cambridge, Massachusetts: MIT Press.

Handel, S. & Erickson, M. L. (2001) A rule of thumb: The bandwidth for timbre invariance is one octave. *Music Perception*, 19(1), 121-126.

Heckerman, D., Horvitz, E. & Nathwani, B. (1992) Towards normative expert systems: Part I The Pathfinder Project. *Methods of Information in Medicine*, 31, 90-105.

Heckerman, D., Mamdani, A. & Wellmann, M. P. (1995) Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3), 24-26.

Helmholtz, H. von (1877) *On the sensations of tone as a physiological basis for the theory of music*. Translated by A. Ellis 1885/1954. Dover, New York.

Hermansky, H. (1987) Why is the formant-frequency difference limen asymmetric? *Journal of the Acoustical Society of America*, 81(S1), S18-S18.

Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems*. Ann Arbor, Michigan: University of Michigan Press.

Horner, A., Beauchamp, J. & Haken, L. (1993) Machine Tongues XVI: Genetic Algorithms and Their Application to FM Matching Synthesis. *Computer Music Journal*, 17(4), 17-29.

Horner, A., Beauchamp, J. & Haken, L. (1993) Methods for Multiple Wavetable Synthesis of Musical Instrument Tones. *Journal of the Audio Engineering Society*, 41(5), 336-356.

Horner, A. & Goldberg, D. E. (1991) Genetic algorithms and computer-assisted music composition *Proceedings of the 4th international conference on genetic algorithms* 437-441.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-441.

Hourdin, C., Charbonneau, G. & Moussa, T. (1997) A Multidimensional Scaling Analysis of Musical Instruments' Time Varying Spectra. *Computer Music Journal*, 21(2), 40-55.

Hourdin, C., Charbonneau, G. & Moussa, T. (1997) A Sound Synthesis Technique Based on Multidimensional Scaling of Spectra. *Computer Music Journal*, 21(2), 40-55.

Howard, D., Disley, A. & Hunt, A. (2007) Timbral adjectives for the control of a music synthesizer. *19th International Congress on Acoustics*, Madrid.

Hunt, A., Wanderley, M. M. & Kirk, R. (2000) Towards a Model for Instrumental Mapping in Expert Musical Interaction. *Proceedings of the 2000 International Computer Music Conference* 209-212.

Hutchins, E. L., Hollan, J. D. & Norman, D. A. (1986) Direct Manipulation Interfaces. In: D. A. Norman & S. W. Draper, ed. *Direct Manipulation Interfaces in User Centered System Design - New Perspectives of Human Computer Interaction*. Lawrence Erlbaum Associates.

Iverson, P. & Krumhansl, C. L. (1993) Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, 94(5), 2595-2603.

Jaffe, D. A. (1995) Ten Criteria for Evaluating Synthesis Techniques. *Computer Music Journal*, 19(1), 76-87.

Johnson, C. G. (1999) Exploring the sound-space of synthesis algorithms using interactive genetic algorithms. *AISB'99 Symposium on Musical Creativity*, Edinburgh, 20-27.

Johnson, C. G. (2003) Exploring sound-space with interactive genetic algorithms. *Leonardo*, 36(1), 51-54.

Kaminskyj, I. (1999) Multidimensional scaling analysis of musical instrument sounds' spectra. *Australasian Computer Music Conference (ACMC)*, Wellington, NZ, 36-39.

Karkoschka, E. (1966) *Notation in New Music*. London: Universal Edition.

Kashino, K. & Murase, H. (1998) Music recognition using note transition context. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, USA, 3593-3596.

Kashino, K., Nakadai, K., Kinoshita, T. & Tanaka, H. (1998) Application of the Bayesian probability network to music scene analysis. In: D. F. Rosenthal &

H. G. Okuno, ed. *Computational auditory scene analysis: Proceedings of the Ijcai-95 Workshop*. Lawrence Erlbaum Associates.

Kendall, R. & Carterette, E. C. (1993) Verbal attributes of simultaneous instrument timbres: I. von Bismarck adjectives. *Music Perception*, 10(4), 445-467.

Kendall, R. & Carterette, E. C. (1993) Verbal attributes of simultaneous instrument timbres: II. Adjectives induced from Piston's orchestration. *Music Perception*, 10(4), 469-502.

Kendall, R. A. & Carterette, E. C. (1991) Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, 8(4), 369-404.

Kendall, R. A., Carterette, E. C. & Hajda, J. M. (1995) Perceptual and acoustical attributes of natural and emulated orchestral instrument timbres. *Proceedings of the International Symposium on Musical Acoustics*, Le Normont, Dourdan, France, 595-601.

Kewley-Port, D. & Watson, C. S. (1994) Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America*, 95(1), 485-496.

Kiger, J. I. (1984) The depth/breadth trade-off in the design of menu driven user interfaces. *International Journal of Man-Machine Studies*, 20(2), 201-213.

Klapper, M. (2000) Working with Csound's ADSYN, LPREAD, and LPRESON Opcodes. In: R. Boulanger, ed. *The Csound Book*. Cambridge, Massachusetts: MIT Press.

Köhler, W. (1915) Akustische Untersuchungen. *Zeitschrift für Psychologie mit Zeitschrift fur Angewandte Psychologie und Charakterkunde*, 58, 59-140.

Koza, J. (1992) *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, Massachusetts: MIT Press.

Krimphoff, J., McAdams, S. & Winsberg, S. (1994) Caractérisation du timbre des sons complexes. II. Analyses acoustiques at quantification psychophysique. *Journal de Physique IV. Colloque C5, supplément au Journal de Physique III*, 4, 625-628.

Krumhansl, C. L. (1989) Why is musical timbre so hard to understand? In: S. Nielzen & O. Olsson, ed. *Structure and Perception of Electroacoustic Sound and Music*. Amsterdam: Elsevier (Excerpta Medica 846), 43-53.

Kruskal, J. B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1-27.

Kruskal, J. B. & Wish, M. (1978) *Multidimensional Scaling*. Newbury Park, California: SAGE Publications.

Landauer, T. K. & Nachbar, D. W. (1985) Selection from alphabetic and numeric menu trees using a touch screen: Breadth, depth and width. *Conference of the SIGCHI on human factors in computing systems*, San Francisco, California, 73-78.

Lerdahl, F. (1987) Timbral hierarchies. *Contemporary Music Review*, 2(1), 135-160.

Lerdahl, F. & Jackendoff, R. (1983) *A generative theory of tonal music*. Cambridge, Massachusetts: MIT Press.

Levitt, T., Agosta, J. & Binford, T. (1990) Model-based influence diagrams for machine vision. In: ed. *Uncertainty in Artificial Intelligence 5*. North-Holland, N.Y., 371-388.

Lichte, W. H. (1941) Attributes of complex tones. *Journal of Experimental Psychology*, 28, 455-480.

Licklider, J. C. R. (1951) Basic Correlates of the Auditory Stimulus. In: S. S. Stevens, ed. *Handbook of Experimental Psychology*. New York: Wiley.

Lyzenga, J. & Horst, J. W. (1997) Frequency discrimination of stylized harmonic synthetic vowels with a single formant. *Journal of the Acoustical Society of America*, 102(3), 1755-1767.

Macdonald, H. (2002) *Berlioz's Orchestration Treatise - a Translation and Commentary*. Cambridge: Cambridge University Press.

Mandelis, J. (2001) Genophone: An Evolutionary Approach to Sound Synthesis and Performance. In: E. Bilotta, E. R. Miranda, P. Pantano & P. Todd, ed. *Proceedings of ALMMA 2002: Workshop on artificial life models for musical applications*. Editoriale Bios, Cosenza, Italy, 108-119.

Mandelis, J. (2002) Adaptive Hyperinstruments: Applying Evolutionary Techniques to Sound Synthesis and Performance. *Proceedings of NIME 2002: New Interfaces for Musical Expression*, Dublin, Ireland, 1-2.

Mandelis, J. & Husbands, P. (2006) Genophone: evolving sounds and integral performance parameter mappings. *International Journal on Artificial Intelligence Tools*, 20(10), 1-23.

Manzolli, J., Maia Jr, A., Fomari, J. & Damiani, F. (2001) The Evolutionary Sound Synthesis Method. *Proceedings of the 9th ACM international conference on Multimedia*, Ottawa, Canada, 585-587.

Manzolli, J., Moroni, A., von Zuben, F. & Gudwin, R. (1999) An Evolutionary Approach Applied to Algorithmic Composition. *Proceedings of VI Brazilian Symposium on Computer Music*, Rio de Janiero, 201-210.

Marozeau, J., de Cheveigne, A., McAdams, S. & Winsberg, S. (2003) The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, 114(5), 2946 – 2957.

Martins, J. M., Pereira, F. C., Miranda, E. R. & Cardoso, A. (2004) Enhancing Sound Design with Conceptual Blending of Sound Descriptors. *Proceedings of the workshop on computational creativity (CC'04)*, Madrid, Spain, 243-255.

Mathes, R. C. & Miller, R. L. (1947) Phase effects in monaural perception. *Journal of the Acoustical Society of America*, 19(5), 780-797.

McAdams, S. (1999) Perspectives on the Contribution of Timbre to Musical Structure. *Computer Music Journal*, 23(3), 85-102.

McAdams, S., Beauchamp, J. & Meneguzzi, S. (1999) Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, 105(2), 882-897.

McAdams, S. & Cunible, J. C. (1992) Perception of Timbral Analogies. *Philosophical Transactions of the Royal Society of London - Series B - Biological Sciences*, 383-389.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G. & Krimphoff, J. (1995) Perceptual scaling of synthesized musical timbres: common dimensions specificities and latent subject classes. *Psychological Research*, 58(3), 177-192.

McDermott, J. (2008) *Evolutionary Computation Applied to the Control of Sound Synthesis*. Ph.D thesis. University of Limerick.

McDermott, J., Griffith, N. J. L. & O'Neill, M. (2005) Toward User-Directed Evolution of Sound Synthesis Parameters. In: *Applications on Evolutionary Computing*. Berlin/ Heidelberg: Springer. 3449/2005.

McDermott, J., Griffith, N. J. L. & O'Neill, M. (2007) Evolutionary GUIs for Sound Synthesis. In: *Applications of Evolutionary Computing*. Berlin/Heidelberg: Springer 547-556.

Mermelstein, P. (1978) Difference limens for formant frequencies of steady-state and consonant-bound vowels. *Journal of the Acoustical Society of America*, 63(2), 572-580.

Miller, J. R. & Carterette, E. C. (1975) Perceptual space for musical structures. *Journal of the Acoustical Society of America*, 58(3), 711-720.

Mintz, D. (2007) *Toward timbral synthesis: A new method for synthesizing sound based on timbre description schemes.* M.S. thesis. University of California, Santa Barbara.

Miranda, E. R. (1995) An Artificial Intelligence Approach to Sound Design. *Computer Music Journal*, 19(2), 59-75.

Miranda, E. R. (1998) Striking the right note with ARTIST: an AI-based synthesiser. In: M. Chemillier & F. Pachet, ed. *Recherches et applications en informatique musicale*. Paris: Editions Hermes, 227- 239.

Miranda, E. R. (2002) *Computer Sound Design*. Oxford: Focal Press.

Mitchell, T. J. & Pipe, A. G. (2005) Convergence Synthesis of Dynamic Frequency Modulation Tones Using an Evolution Strategy. In: *Applications on Evolutionary Computing*. Berlin / Heidelberg: Springer. 3449/2005.

Moorer, J. A. (1973). *The Heterodyne Filter as a Tool for Analysis of Transient Waveforms*. Memo 208, Stanford Artificial Intelligence Laboratory, Stanford, California.

Moorer, J. A. (1975). *On the loudness of complex, time variant tones*. Report no STAN-M-4, Center for Computer Research in Music and Acoustics, Stanford University.

Moravec, O. & Stepánek, J. (2003) Verbal description of musical sound timbre in Czech language. *Proceedings of the Stockholm Music Acoustics conference (SMAC'03)*, Stockholm, 643-645.

Moroni, A., Manzolli, J., von Zuben, F. & Gudwin, R. (2000) Vox Populi: an interactive evolutionary system for algorithmic music composition. *Leonardo Music Journal*, 10, 49-55.

Nadi, F., Agogino, A. & Hodges, D. (1991) Use of influence diagrams and neural networks in modeling semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, 4(1), 52-58.

Native Instruments (2009) *Reaktor*. [CD-ROM] Mac OS X. Berlin: Native Instruments.

Nefian, A. V., Liang, L., Pi, X., Liu, X. & Murphy, K. (2002) Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *EURASIP Journal on Applied Signal Processing*, 11, 1274–1288.

Nicol, C. A. (2005) *Development and Exploration of a Timbre Space Representation of Audio*. Ph.D thesis. University of Glasgow.

Nielsen, J. (1994) Heuristic Evaluation. In: J. Nielsen & R. Mack, ed. *Usability Inspection Methods*. New York: Wiley, 25-62.

Nielsen, J. (1994) *Usability engineering*. San Francisco: Morgan Kaufmann.

Norman, D. A. (1988) *The Psychology of Everyday Things*. New York: Basic Books.

Norman, K. L. & Chin, J. P. (1988) The effect of tree structure on search in a hierarchical menu selection system. *Behaviour and Information Technology*, 7(1), 51-65.

Nykänen, A. & Johannsen, Ö. (2003) Development of a language for specifying saxophone timbre. *Proceedings of the Stockholm Music Acoustics Conference (SMAC'03)* 647-650.

Osgood, C. E., Suci, G. H. & Tannenbaum, P. H. (1957) *The measurement of meaning*. Urbana, Illinois: University of Illinois Press.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan-Kaufmann.

Pereira, F. C. & Cardoso, A. (2003) Optimality principles for conceptual blending: a first computational approach. *Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, 1(4), 351-369.

Piston, W. (1955) *Orchestration*. New York: W W Norton.

Plomp, R. (1970) Timbre as a Multidimensional Attribute of Complex Tones. In: R. Plomp & G. F. Smoorenberg, ed. *Frequency Analysis and Periodicity Detection in Hearing*. Leiden: Suithoff, 397-414.

Plomp, R. (1976) *Aspects of tone sensation*. New York: Academic Press.

Plomp, R. & Steeneken, J. M. (1969) Effect of phase on the timbre of complex tones. *Journal of the Acoustical Society of America*, 46(2B), 409-421.

Plomp, R. & Steeneken, J. M. (1971) Pitch versus timbre. *Proceedings of the 7th International Congress of Acoustics*, Budapest, 377-380.

Polfreman, R. & Sapsford-Francis, J. (1995) A Human-Factors Approach to Computer Music Systems User-Interface Design. *Proceedings of the 1995 International Computer Music Conference*, Banff, Canada.

Pratt, R. L. & Doak, P. E. (1976) A subjective rating scale for timbre. *Journal of Sound and Vibration*, 45(3), 317-328.

Pressing, J. (1992) *Synthesiser Performance and Real-Time Techniques*. Madison, Wisconsin: A-R Editions.

Raphael, C. (2002) A Bayesian Network for Real-Time Musical Accompaniment. In: *Advances in Neural Information Processing Systems, NIPS 14*. MIT Press.

REALSoftware (2006) *REALbasic*. [CD-ROM] Mac OS X. Austin, TX: REAL Software, Inc.

Richardson, E. G. (1954) The transient tones of wind instruments. *Journal of the Acoustical Society of America*, 26(6), 960-962.

Riionheimo, J. & Välimäki, V. (2003) Parameter Estimation of a Plucked String Synthesis Model Using a Genetic Algorithm with Perceptual Fitness Calculation *EURASIP Journal on Applied Signal Processing*, 2003(8), 791-805.

Rimsky-Korsakov, N. (1891) *Principles of Orchestration*. Translated by E. Agate, 1922. New York: E. F. Kalmus.

Risset, J. C. & Wessel, D. L. (1999) Exploration of Timbre by Analysis and Synthesis. In: D. Deutsch, ed. *The Psychology of Music*. San Diego: Academic Press.

Roads, C. (1988) Introduction to granular synthesis. *Computer Music Journal*, 12(2), 11-13.

Roads, C. (1996) *The Computer Music Tutorial*. Cambridge, Massachusetts: MIT Press.

Rodet, X. (1984) Time-domain formant-wave-function synthesis. *Computer Music Journal*, 8(3), 9-14.

Rodet, X., Potard, Y. & Barriere, J. B. (1984) The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General. *Computer Music Journal*, 8(3), 15-31.

Rolland, P.-Y. & Pachet, F. (1996) A Framework for Representing Knowledge about Synthesizer Programming. *Computer Music Journal*, 20(3), 47-58.

Ruffner, J. W. & Coker, G. W. (1990) A Comparative Evaluation of the Electronic Keyboard Synthesizer User Interface. *Proceedings of the Human Factors Society 34th Annual Meeting* 477-481(5).

Saldanha, E. L. & Corso, J. F. (1964) Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, 36(11), 2021-2026.

Sandell, G. (1989) Effect of spectrum and attack properties on the evaluation of concurrently sounding timbres. *Program of the 118th meeting of the Acoustical Society of America* S59-S59.

Sandell, G. (1989) Perception of concurrent timbres and implications for orchestration. *Proceedings of the International Computer Music Conference (ICMC 1989)*, Columbus, Ohio, 268-72.

Sandell, G. & Martens, W. (1995) Perceptual evaluation of principal components-based synthesis of musical timbres. *Journal of the Audio Engineering Society*, 43(12), 1013-1028.

Schaeffer, P. (1966) *Traité des objets musicaux*. Paris: Ed. du Seuil.

Schatter, G., Züger, E. & Nitschke, C. (2005) A synaesthetic approach for a synthesizer interface based on genetic algorithms and fuzzy sets. *Proceedings of the International Computer Music Conference, 2005 (ICMC 2005)*, Barcelona, Spain, 664-667.

Schönberg, A. (1911) *Theory of Harmony*. Translated by R. E. Carter, 1978. London: Faber & Faber.

Seago, A. (2004). *Analysis of the synthesizer user interface: cognitive walkthrough and user tests.* Technical report TR2004/15, Department of Computing, Open University.

Seashore, C. E. (1967) *The psychology of music*. New York: Dover, New York.

Sethares, W. (2005) *Tuning, Timbre, Spectrum, Scale*. Berlin, Heidelberg, New York: Springer-Verlag.

Shepard, R. N. (1982) Geometrical approximations to the structure of musical pitch. *Psychological Review*, 89 (4), 305-333.

Shneiderman, B. (1983) Direct manipulation: A step beyond programming languages. *IEEE Computer*, 16(8), 57-69.

Shneiderman, B. (1997) *Designing the User-Interface: Strategies for Effective Human-Computer Interaction*. Reading, MA: Addison-Wesley.

Slawson, A. W. (1989) Sound Structure and Musical Structure: The Role of Sound Color. *Proceedings of the Marcus Wallenberg symposium,1988*, Lund, Sweden

Slawson, W. (1968) Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America*, 43(1), 87-101.

Slawson, W. (1985) *Sound Color*. Berkeley: University of California Press.

Smalley, D. (1986) Spectromorphology and structuring processes. In: S. Emmerson, ed. *The language of electroacoustic music*. New York: Harwood Academic Publishers, 61-93.

Smalley, D. (1997) Spectromorphology: explaining sound-shapes. *Organised Sound*, 2(2), 107-126.

Smith, J. O. (1991) Viewpoints on the history of digital synthesis. *Proceedings of the International Computer Music Conference*, Montreal, Canada, 1-10.

Smith, J. O. I. (1992) Physical modeling using digital waveguides. *Computer Music Journal*, 16(4), 74-91.

Solomon, L. N. (1958) Semantic Approach to the Perception of Complex Sounds. *Journal of the Acoustical Society of America*, 30(5), 421-425.

Solomon, L. N. (1959) Search for physical correlates to psychological dimensions of sounds. *Journal of the Acoustical Society of America*, 31(4), 492-497.

Stumpf, C. (1926) *Die Sprachlaute*. Berlin & New York: Springer Verlag.

Takagi, H. (2001) Interactive Evolutionary Computation: Fusion of the Capabilities of EC Optimization and Human Evaluation *Proceedings of the IEEE* 1275–1296.

Takala, T., Hahn, J., Gritz, L., Geigel, J. & Lee, J. W. (1993) Using physically-based models and genetic algorithms for functional composition of sound signals, synchronized to animated motion. *Proceedings of the International Computer Conference (ICMC'93)*, Tokyo, Japan, 180-185.

The Mathworks (2007) *MATLAB*. [CD-ROM] Mac OS X. Natick, Massachusetts: The MathWorks.

Thimbleby, H. (2001) The Computer Science of Everyday Things. *Proceedings of the Australasian User Interface Conference*, Goldcoast: Washington, 3-11.

Toiviainen, P., Kaipainen, M. & Louhivuori, J. (1995) Musical timbre: similarity ratings correlate with computational feature space distances. *Journal of New Music Research*, 24(3), 282-298.

Tolonen, T., Välimäki, V. & Karjalainen, M. (1998). *Report no 48: Evaluation of Modern Sound Synthesis Methods*. Helsinki University of Technology, Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing.

Truax, B. (1980) The inverse relation between generality and strength in computer music programs. *Journal of New Music Research*, 9(1), 49-57.

Truax, B. (1988) Real-time granular synthesis with a digital signal processor. *Computer Music Journal*, 12(2), 14-26.

Vercoe, B. (1985). *The Csound Music Synthesis Language*. Boston, MA, Media Laboratory, Massachusetts Institute of Technology.

Vertegaal, R. (1994) *An evaluation of input devices for timbre space navigation*. M. Phil dissertation. University of Bradford.

Vertegaal, R. & Bonis, E. (1994) ISEE: An Intuitive Sound Editing Environment. *Computer Music Journal*, 18(2), 21-29.

Vertegaal, R. & Eaglestone, B. (1996) Comparison of input devices in an ISEE direct timbre manipulation task. In: *Interacting with Computers*. Butterworth-Heinemann.

Wanderley, M. M. & Orio, N. (2002) Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI. *Computer Music Journal*, 26(3), 62-76.

Wedin, L. & Goude, G. (1972) Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology*, 13(3), 228-240.

Weidenaar, R. (1995) *Music from the Telharmonium*. Metuchen, NJ: Scarecrow Press.

Wessel, D. L. & Wright, M. (2002) Problems and Prospects for Intimate Musical Control of Computers. *Computer Music Journal*, 26(3), 11-22.

Wharton, C., Rieman, J., Lewis, C. & Poison, P. (1994) The cognitive walkthrough method: A practitioner's guide. In: J. Nielsen & R. L. Mack, ed. *Usability Inspection Methods*. New York: John Wiley & Sons, 105-140.

Williges, R. C., Williges, B. H. & Elkerton, J. (1987) Software Interface Design. In: G. Salvendy, ed. *Handbook of Human Factors*. New York: John Wiley & Sons.

Winsberg, S. & Carroll, J. D. (1988) A Quasi-Nonmetric Method for Multidimensional Scaling via an Extended Euclidian Model. *Psychometrika*, 53, 217-229.

Wishart, T. (1986) Sound Symbols and Landscapes. In: S. Emmerson, ed. *The Language of Electroacoustic Music*. Basingstoke: Macmillan Press.

Wishart, T. (1996) *On Sonic Art*. Amsterdam: Overseas Publishers Association B.V.

Xenakis, I. (1971) *Formalized Music*. Bloomington and London: Indiana University Press.

Appendix I - Design and implementation of search software

Construction of attribute spaces

Formant space

All the 1690 sounds contain 73 harmonics with a base frequency of 110 Hz, and each has three prominent formants. The first attribute, the centre frequency of formant I, varies between 110 and 440 Hz; the second attribute, the centre frequency of formant II, varies between 550 and 2200; the third attribute, the centre frequency of formant III, varies between 2200 and 6600 Hz.

In order to populate this space, 1690 AIFF files were created, each one generated by a separate Csound 'instrument' in the .orc file. A typical instrument listing is given here.

```
instr      930                              ;        I_8_II_2_III_10.aif
kenv       linen          30000,     0.4,     p3,      0.4
asig       oscil          kenv,      110,     1
aout1      pareq asig,    243.175    ,        10   ,   6   ,   0
aout2      pareq asig,    616        ,        10   ,   6   ,   0
aout3      pareq asig,    6100.773   ,        10   ,   6   ,   0
afin = aout1 +aout2 + aout3
aout balance afin, asig
outs aout*0.9, aout*0.9
endin
```

The audio signal **asig** is generated by the Csound opcode **oscil**, with a base frequency of 110 Hz, and using a wavetable whose spectrum is defined in function table 1 (listed in the score file). The signal **asig** is given an amplitude envelope **kenv** with attack and decay times of 0.4 seconds, and split into three audio streams, each of which is input to a peaking filter **pareq** (providing the three formant peaks characterising the stimuli in this space). The outputs **aout1**, **aout2** and **aout3** from these filters have (for this instrument) peak centre frequencies of 243.175, 616 and 6100.773 Hz respectively, and are summed to produce the signal **afin**; which, in turn, is scaled to prevent clipping, to produce the instrument output **aout**.

The accompanying score file contained the function table specification – a waveform of 73 harmonics of equal amplitude - and 'played' the instruments to an output AIFF file. A section of the file is reproduced here.

```
f1.0 512.10.1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

i     1     3     2     ; audiofile     I_1_II_1_III_1.aif
i     2     6     2     ; audiofile     I_1_II_1_III_2.aif
i     3     9     2     ; audiofile     I_1_II_1_III_3.aif
i     4     12    2     ; audiofile     I_1_II_1_III_4.aif
etc
```

The resulting audio file was submitted to spectrum analysis to verify its spectrum, and then split into 1690 files using an audio editor. Each file was then normalised to a level –3.09 dBfs short of full amplitude.

SCG-EHA space

All the 1815 sounds contain 20 harmonics with a base frequency of 311 Hz; the variable attributes of the sounds which form each of the three axes are :

- Rise time, ranging from 0.01 to 0.2 seconds in 11 logarithmic steps. In all cases, the attack envelope was linear.
- EHA.- attenuation of even harmonics relative to the odd ones in the range 0 dB to 10 dB – again in 11 steps.
- SCG - spectral centroid, in the range 3.000 to 8.000, in 15 linear steps. This corresponds to a spectral centroid range of 933 Hz to 2488 Hz.

The 1815 sound files were prepared using Csound. The extract from the 'orchestra' file shows an instrument which generates a stimulus with an attack time of 0.060 seconds (**irisetime**), and a harmonic structure with a base frequency of 311 Hz; the harmonic amplitudes are given by parameters **p4** to **p23**, the values of which appear in the 'score' file.

```
instr 7

        irisetime = 0.060

        aharm1          oscil   p4      ,       311     ,       1
        aharm2          oscil   p5      ,       622     ,       1
        aharm3          oscil   p6      ,       933     ,       1
        aharm4          oscil   p7      ,       1244    ,       1
        aharm5          oscil   p8      ,       1555    ,       1
        aharm6          oscil   p9      ,       1866    ,       1
        aharm7          oscil   p10     ,       2177    ,       1
        aharm8          oscil   p11     ,       2488    ,       1
        aharm9          oscil   p12     ,       2799    ,       1
        aharm10 oscil   p13     ,       3110    ,       1
        aharm11 oscil   p14     ,       3421    ,       1
        aharm12 oscil   p15     ,       3732    ,       1
        aharm13 oscil   p16     ,       4043    ,       1
        aharm14 oscil   p17     ,       4354    ,       1
        aharm15 oscil   p18     ,       4665    ,       1
        aharm16 oscil   p19     ,       4976    ,       1
        aharm17 oscil   p20     ,       5287    ,       1
        aharm18 oscil   p21     ,       5598    ,       1
        aharm19 oscil   p22     ,       5909    ,       1
        aharm20 oscil   p23     ,       6220    ,       1

        aenv envlpx 15, irisetime, 0.6, 0.2, 3 , 1, 0.01

        asig = (aharm1 + aharm2 + aharm3 + aharm4 + aharm5 + aharm6 + aharm7 + aharm8 + aharm9 +
        aharm10 + aharm11 + aharm12 + aharm13 + aharm14 + aharm15 + aharm16 + aharm17 + aharm18 +
        aharm19 + aharm20)/20.

        afin = asig.* aenv

        out afin

endin
```

The following is an extract from the 'score' file. Each of the four calls to **instr 7** shown here gives the harmonic amplitudes for one audio stimulus. The spectral centroids are, in this case, 3.000, 3.396, 3.813 and 4.242 respectively.

```
i7      45 0.6     9000.00 2312.31 1504.17 747.91 654.70 386.46 378.52 241.91 251.39 168.21  181.32     125.00
        138.13     97.25      109.42     78.24      89.24 64.59 74.46 54.41       ; spectral centroid = 3.000

i7      46.5 0.6 9000.00 2531.17 1736.00 896.19 807.67 488.23    487.92    317.31    334.85    227.15    247.92
        172.86     193.03     137.22     155.79     112.35     129.16     94.17 109.34 80.43 ; spectral centroid = 3.396


i7      48 0.6 9000.00 2750.75 1980.67 1058.42 979.77 605.37 616.27 407.25 435.90 299.46 330.60 232.93
        262.62     188.36     215.62     156.70     181.47     133.23     155.68     115.22     ; spectral centroid = 3.813


i7      49.5 0.6 9000.00 2969.51 2236.08 1233.47 1170.29 737.78 763.97 512.35 555.56 386.13 430.79     306.46
        348.59     252.07     290.76     212.82     248.11     183.31     215.48     160.39     ; spectral centroid = 4.242


etc
```

As with the formant space stimuli, the resulting audio file was verified using spectrum analysis, and then split into 1815 files using an audio editor. Each file was then normalised to a level –3.09 dBfs short of full amplitude.

MDS space

Because of the high dimensionality of the MDS space, the generation of static soundfiles using Csound would have been impractical – 823,543 rather than 1,690 or 1,815 files would have been required. Instead, the waveforms were generated dynamically in the software, using PortAudio.

The process of generating the reduced dimensionality space is as described in chapter seven.

Appendix II -Program design

WCL-2 version

Formant space and SCG-EHA space

At initialisation, the program performs the following steps:

1. Initialises the probability table, seeding the 13x13x10 cells (in the case of the formant space) or the 11x11x15 cells (in the case of the SCG-EHA space) with a value of 100.

2. Sets up the target sound coordinates

3. Creates the monitoring files used for analysing the operation of the program

4. Identifies the weighted centroid of the probability table (at the outset, this is located at the centre of the table)

5. Generates two probe sounds A and B. These probes are generated such that the Euclidean distance between them is never below a specified minimum, in order to ensure sufficient timbral dissimilarity.

Each time the subject makes a choice, the following steps are executed:

6. The probability table is updated such that the values of those cells which are closer to the chosen probe are multiplied by √2 – the values of all other cells are divided by √2.

7. The table is then normalised to prevent cell values exceeding computable limits.

8. Two new probe sounds A and B are generated. Again, these probes are generated such that that there is sufficient timbral dissimilarity between them – in addition, the new probes are selected such that a line connecting the pair is more or less orthogonal in at least one plane to a line connecting the previous pair. The purpose of this is to ensure a richer distribution of probabilities in the table, and a progressive shift in the position of the weighted centroid in more than one dimension.

9. Finally, the weighted centroid of the probability table is recalculated . The probe coordinates, the coordinates of the weighted centroid and the Euclidean distance between the weighted centroid and the target are all recorded by the software for later analysis.

MDS space

Broadly speaking, the operation of the algorithm is the same as that for the three-dimensional spaces – the main differences are given here.

At initialisation, the program performs the following steps:

1. Sets up a buffer of 51200*2 bytes (each sample point is 16 bit = 2 bytes),

2. Initialises the transform data

3. Initialises the probability table, seeding the 7^7 cells of the space with a value of 100.

4. The remaining steps are as steps 2- 9 above.

WCL-7 version

Formant space and SCG-EHA space

The initialisation steps 1-4 are the same as those described in section 6.4.1.2. In step 5, seven probes are generated, such that the Euclidean distance between them is never below a specified minimum, in order to ensure sufficient timbral dissimilarity.

Each time the subject makes a choice, the following steps are executed:

10. The value of each probability table cell is multiplied by a factor whose value is in inverse proportion to the Euclidean distance between the target and the selected probe.

11. The table is then normalised to prevent cell values exceeding computable limits.

12. Seven new probe sounds A and B are generated. Again, these probes are generated such that that there is sufficient timbral dissimilarity between them .

13. Finally, the weighted centroid of the probability table is recalculated . The probe coordinates, the coordinates of the weighted centroid and the Euclidean distance between the weighted centroid and the target are all recorded by the software for later analysis.

MDS space

At initialisation, the program performs the following steps:

1. Sets up a buffer of 51200*2 bytes.

2. Initialises the transform data.

3. Initialises the probability table, seeding the 7^7 cells of the space with a value of 100.

4. Sets up the target sound coordinates

5. Creates the monitoring files used for analysing the operation of the program

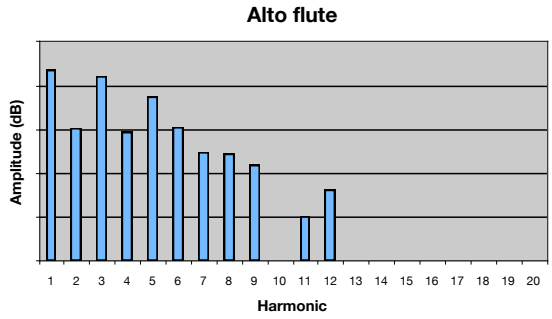
6. Identifies the weighted centroid of the probability table (at the outset, this is located at the centre of the table)

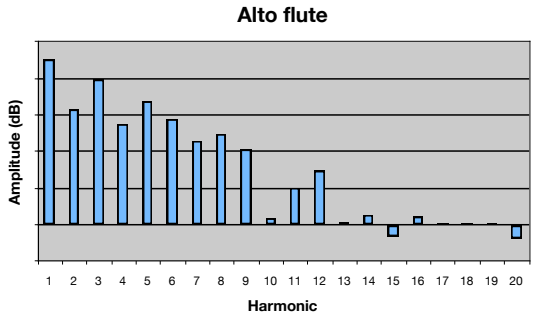
7. Generates seven probes such that the Euclidean distance between them is never below a specified minimum, in order to ensure sufficient timbral dissimilarity.

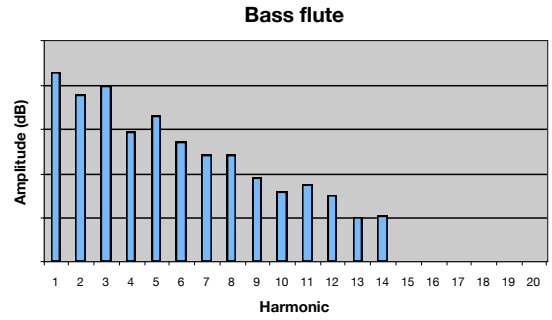
The remaining steps are as steps 10 – 13 above.

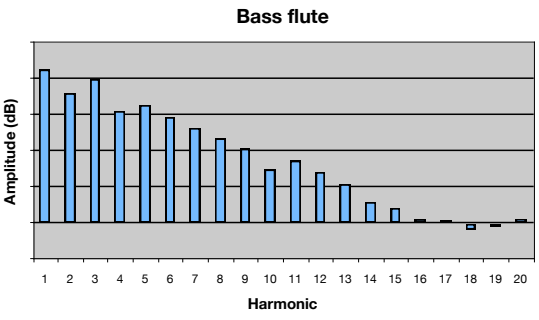
Appendix III - Original and reconstructed heterodyne spectra

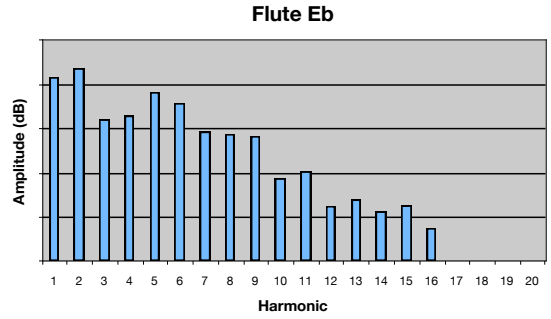
Spectra from H_{LTAS_dB}	Spectra from $H_{LTAS\_dB\_reconstructed}$

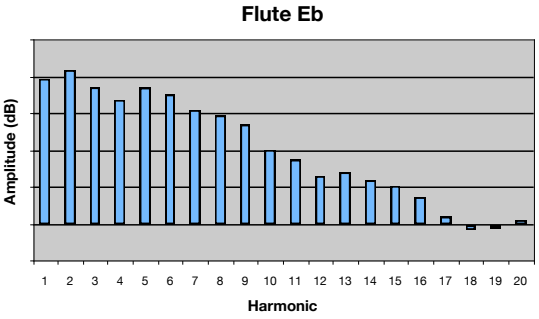




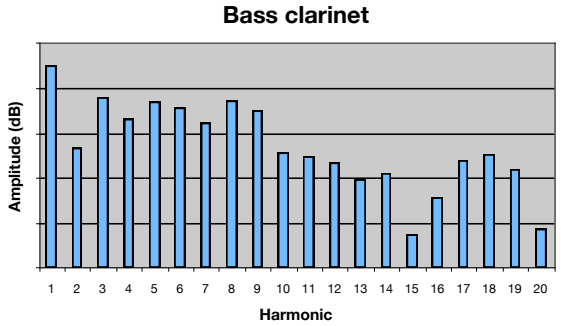


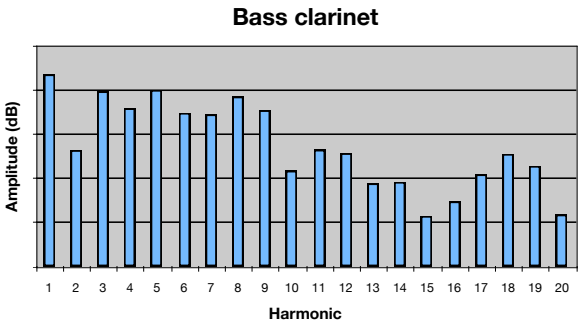


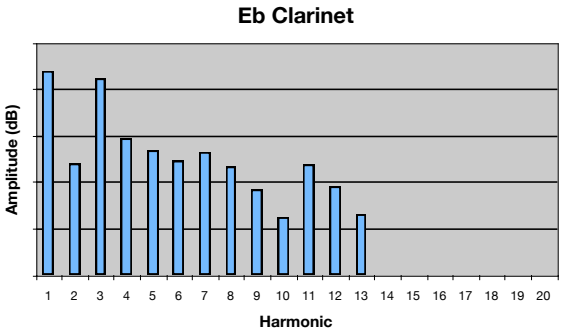


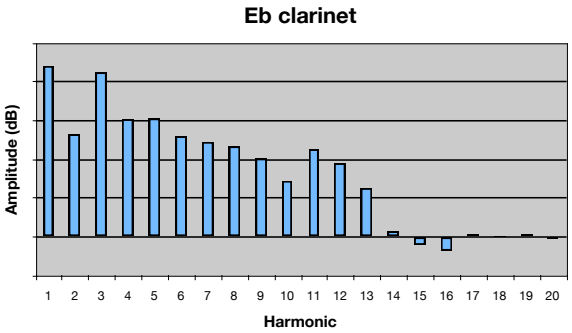


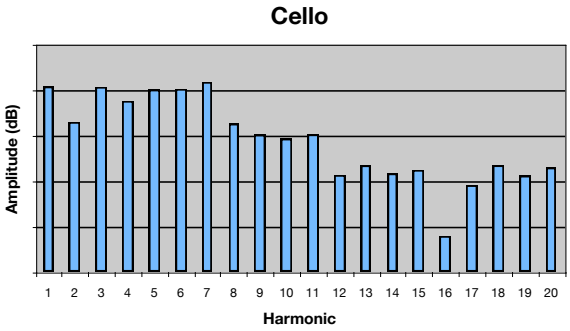
Spectra from H_{LTAS_dB}	Spectra from $H_{LTAS\_dB\_reconstructed}$

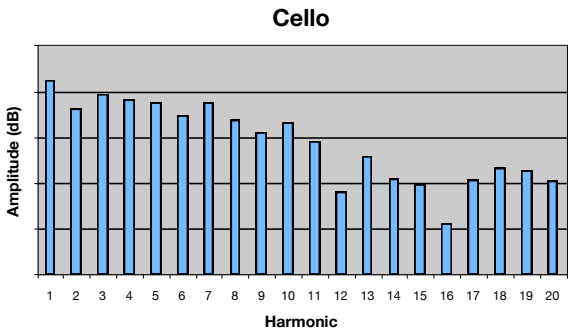




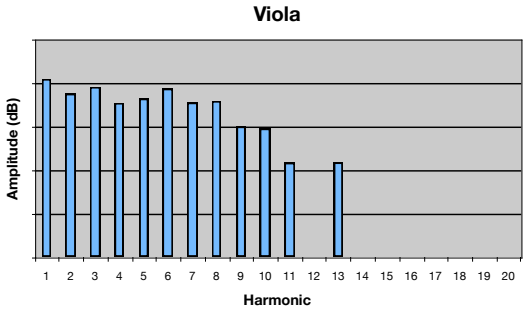


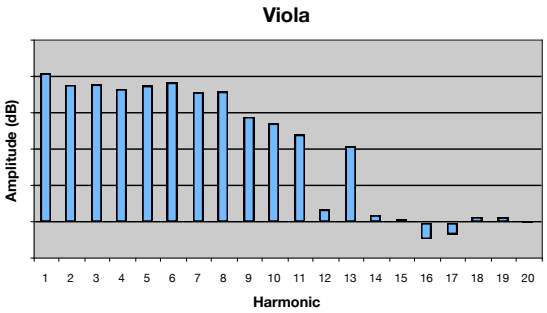


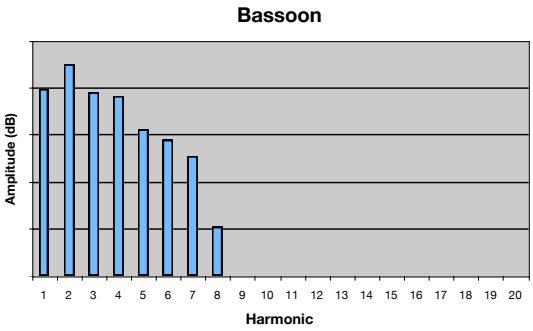


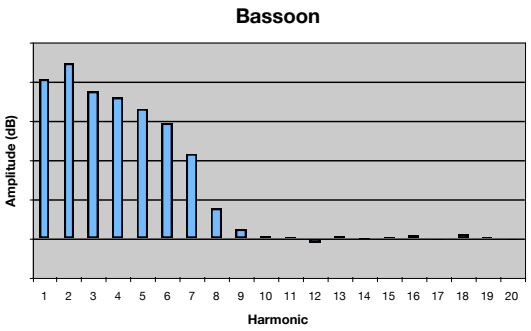


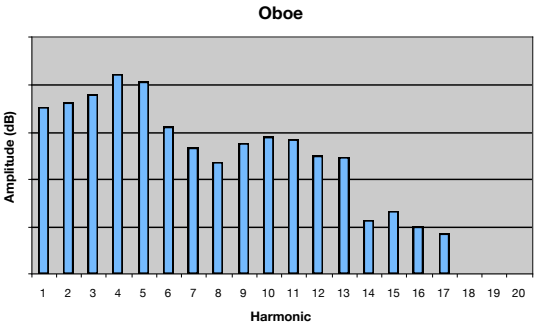
Spectra from H_{LTAS_dB}	Spectra from $H_{LTAS\_dB\_reconstructed}$

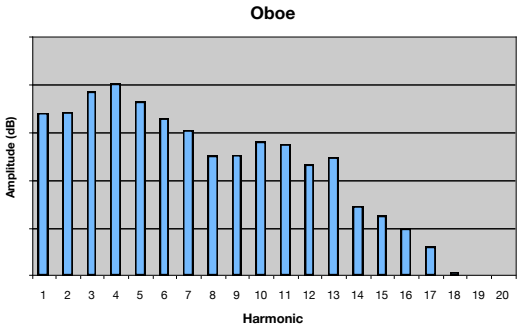




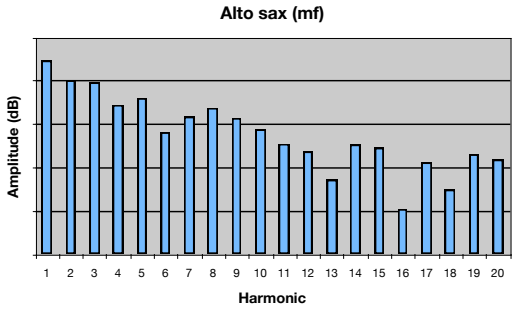


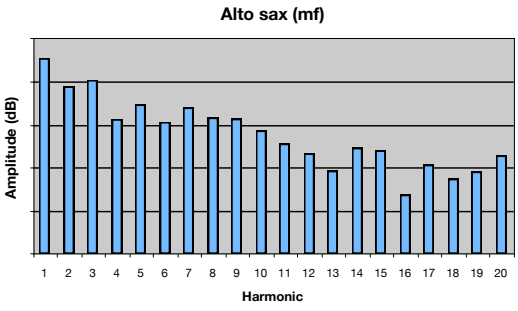


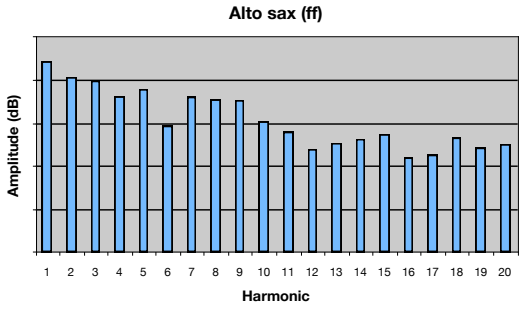


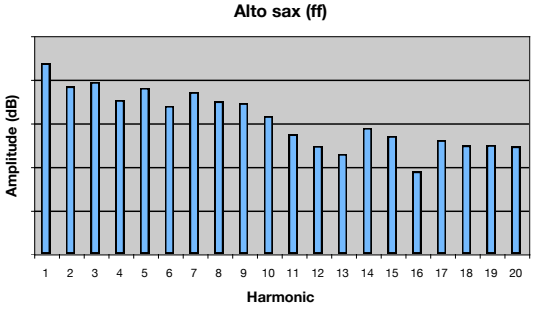


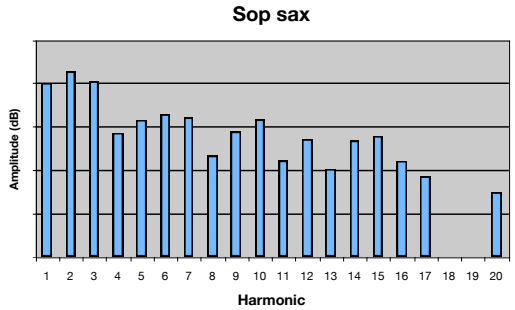
Spectra from H_{LTAS_dB}	Spectra from $H_{LTAS\_dB\_reconstructed}$

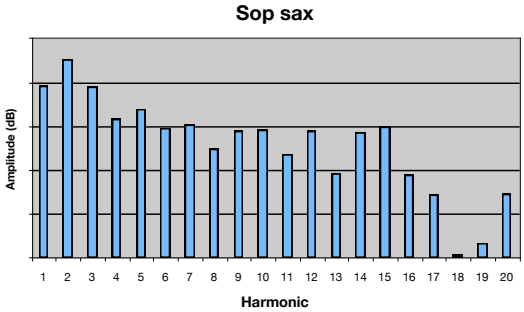




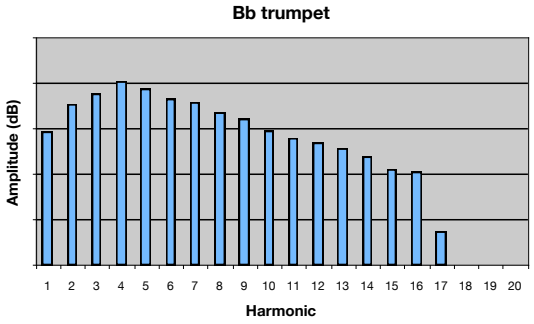


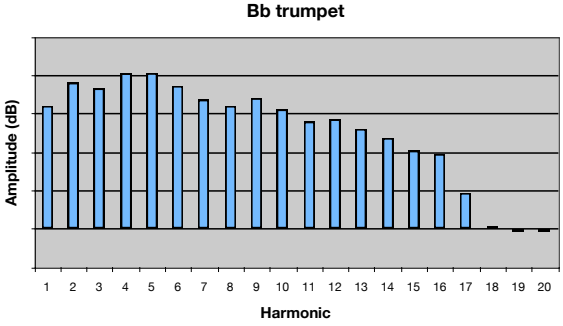


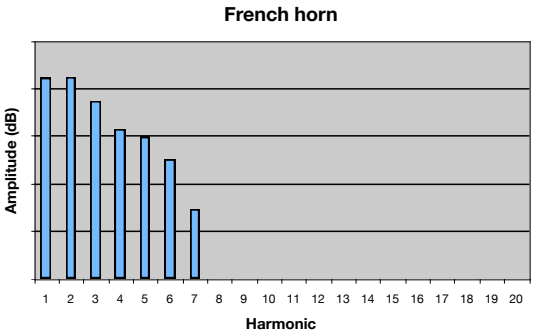


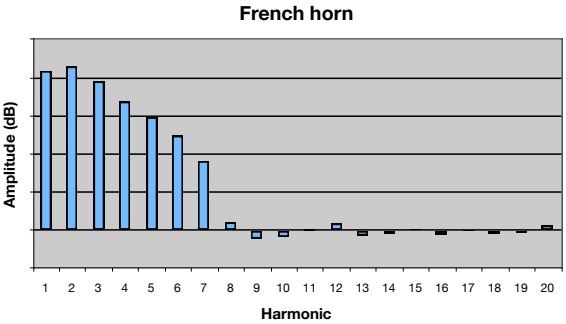


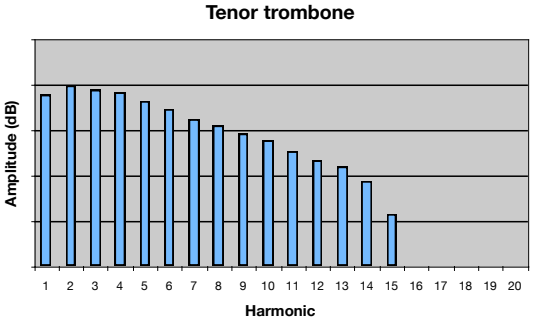
Spectra from H_{LTAS_dB}	Spectra from $H_{LTAS\_dB\_reconstructed}$

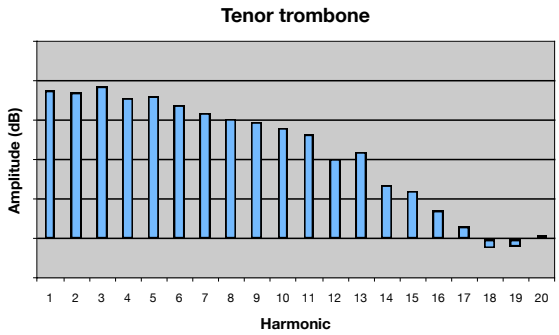












Appendix IV - Research ethics approval application

University Research Ethics Committee
Research Ethics Review Panel

Application for Ethics Approval for a Research Project Involving Human Participants
Staff and PGR MPhil and PhD students

Please complete this form to apply for ethical approval of your research if it involves work with human participants. Please consult the University's Ethics Policy and ensure the completed form is submitted to your RERP at least two weeks before the date of the next meeting. If you subsequently change your project you must notify the RERP (or, if you are a MPhil/PhD student, notify the Research Student Progress Group via the annual monitoring process).

Please expand the form as necessary.

1. Title of project: **Software user testing**

2. Title and name of principal investigator responsible for the project:.**Allan Seago**

Status (e.g., Senior Lecturer; PhD student): **Senior Lecturer**

Department/Research Institute: **Sir John Cass Department of Art Media & Design**
Phone:.**0207 320 2841**
Email:. **a.seago@londonmet.ac.uk**

3. Who will conduct the research? Are they suitably experienced/trained to do so?

Myself. I have successfully run similar tests in previous years.

4. Have you indicated what contribution the research will make and how it will be funded?

The research is cross-disciplinary and is intended to contribute to the fields of human-computer interaction, psychoacoustics and computer music. Application has been made for funding to the SJCAMD RCF budget.

5. When and where will the research be conducted?

May 2008, in room 619 Commercial Road

6. Have you consulted and applied other ethics standards and procedures relevant to your research, including those of other bodies, institutions, professions or regulatory bodies?

Code of Ethics and Conduct of the British Psychological Society. There is also guidance on the conduct of psychoacoustical tests in Cook P, Music, Cognition, and Computerized Sound: An Introduction to PsychoAcoustics (MIT Press).
.
7. Aims of the research, including any hypothesis to be used.

To compare the performance of three software programs, each employing a

different user-driven search strategy designed to enable a user to select a sound out of a search space of possible candidates. Each strategy will be employed on two distinctly different search spaces, making six tests in all.

8. Description of the procedures to be used (give sufficient detail for the group to be clear about what is involved in the research). Please append to the application any instructional leaflets, letters, questionnaires, forms or other documents which will be issued to participants.

In order to investigate this, a set of user tests is to be conducted, using undergraduate students from the music subject area in the Sir John Cass Department of Art Media and Design. Each subject will run all six tests. In each test, the subject will be presented with a target sound, and asked to manipulate software controls designed to move a probe sound through the search space. The purpose is to make the probe sound aurally indistinguishable from the target sound. Each program will log the subject's actions for later analysis.

9.1. How will you obtain freely given, explicit and informed consent, preferably in writing, before the research begins? Please attach a consent form and information sheet as appropriate.

A clear description of the nature of the tests and how long they will last will be given, both in email communication and verbally before the tests begin.

9.2. If not, how do you justify this in the proposal?

9.3. Will you renegotiate consent throughout the life of the research?

Not applicable.

9.4. Have you given or will you give advance information, in writing about the project, and informed respondents on all points in the guidelines?

Yes.

9.5. Have you gained or will you gain informed consent for the use of tape recording and other data collection methods?

Not applicable.

10. Approximate number of participants. Also include nature of participants (e.g., University students, primary school children).

Twenty undergraduate students

11. Have you considered issues relating to gatekeepers and consent with the very young, the very old or respondents who may be ill or who are mentally vulnerable or impaired or others who may find it difficult to deny consent?

Not applicable

12.1. Have you been or will you be open and honest with participants about the research, its purpose and application?

Yes.

12.2. If not, how do you justify withholding information from them?

Not applicable

13. Have you assured participants that they can withdraw from the research at any time without penalty?

Yes.

14.1. Have you ensured or will you ensure confidentiality and anonymity of participants' identity and data?

Yes. Some form of identification of individual subjects is necessary in order to compare individual performance with different software. However, care will be taken to preserve confidentiality.

14.2. How will you work on store data in encoded form?

Question not entirely clear. However, the raw data will be accessed and analysed only by myself.

14.3. How will you ensure that the data be available only to specified researchers, for the purpose for which it was collected, and not used more widely without consent?

All data will be kept securely on researcher's laptop (password protected), and backed up onto securely stored media. All other data on department computers will be deleted.

14.4. Have you complied with the requirements of the Data Protection Act?

Yes.

15.1. Have you assessed any risks to participants and researchers involved in the research (e.g., physical and mental discomfort or danger, impact on the individual)?

The stimuli presented in the tests can be tiring to listen to over time. As in all work involving audio equipment, there is always the risk of damage to hearing due to accidental exposure to excessive sound levels.

15.2. How will you ensure they are protected from harm?

Breaks will be made part of the testing schedule. Care will be taken to ensure that headphone amplifiers are set at an initial low level; participants will be advised that they can adjust the volume themselves to a comfortable level .

15.3. How will you inform them in advance of any risks and protective procedures involved?

Participants will be informed of the above.

15.4. Might medical care or aftercare be required?

No.

15.5. Will there be administration of drugs (including caffeine, alcohol) and if yes, why?

No

16.1. Have you given full information about the purpose of the research and what its outcomes will be?

Yes.

16.2. Will you inform participants of the actual outcomes?

Results are briefly discussed with participants immediately after the test. Participants are in addition free (indeed welcome) to communicate with me on any aspect of the tests.

16.3. Have you told participants if you intend to do any of these things, and can you honour this intention?

Yes.

17. Will the participants be paid? If yes, please state if fee, expenses, honorarium and give details of the reason form payment.

Participants will be paid £10. The tests can be quite tedious, and it is, in general, the practice to pay participants for this kind of work.

18. If your research involves a client group or other co-researchers, have you clarified ethical issues and ownership of data, access and rights to publish, and agreed them in writing in advance?

Yes.

19. How will you advise participants as to dissemination of research findings, and what information will you send them?

Emails will be sent to all participants inviting them to research seminars in which the findings and conclusions of the work will be presented.

20. Are there any other matters which you consider relevant to the consideration of this proposal ? If so, please elaborate below:

No.

21. Declaration

You should bear in mind that this is only one part of the ongoing process of conducting research in an ethically sound manner. You should always keep your project under review for ethics implications.

Approval is given on the basis of the submitted proposal. If there are substantial changes to the project in the future, you must reapply to the RERG.

I undertake to abide by the accepted ethical principles, the University's Code of Good Research Practice and any other appropriate professional code(s) of practice in carrying out this project.

Personal data will be treated in the strictest confidence and not passed on to others without the written consent of the participant.

The nature of the investigation and any possible risks will be fully explained to intending participants, and they will be informed that:

they are in no way obliged to volunteer if there is any personal reason (which they are under no obligation to divulge) why they should not participate in the programme and they may withdraw from the programme at any time, without disadvantage to themselves and

without being obliged to give any reason.

Name of principal investigator. **Allan Seago**

Signed. . . . . Date.

ethlmu17.doc
10 August 2004